From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management

Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, Somayajulu Sripada

From Data to Text in the Neonatal Intensive Care Unit: Using NLG Technology for Decision Support and Information Management

Albert Gatt [a,*] François Portet [b] Ehud Reiter [a]
Jim Hunter [a] Saad Mahamood [a] Wendy Moncur [a]
Somayajulu Sripada [a]

[a] *Department of Computing Science*
University of Aberdeen
King's College
Aberdeen AB24 3UE, UK
E-mail: {a.gatt,e.reiter,j.hunter,s.mahamood, w.moncur,s.sripada}@abdn.ac.uk
[b] *Laboratoire d'Informatique de Grenoble*
Equipe GETALP Bâtiment IMAG B
385 Avenue de la Bibliothque 38400
Saint Martin d'Hères, France
E-mail: francois.portet@imag.fr

Contemporary Neonatal Intensive Care Units collect vast amounts of patient data in various formats, making efficient processing of information by medical professionals difficult. Moreover, different stakeholders in the neonatal scenario, which include parents as well as staff occupying different roles, have different information requirements. This paper describes recent and ongoing work on building systems that automatically generate textual summaries of neonatal data. Our evaluation results show that the technology is viable and comparable in its effectiveness for decision support to existing presentation modalities. We discuss the lessons learned so far, as well as the major challenges involved in extending current technology to deal with a broader range of data types, and to improve the textual output in the form of more coherent summaries.

Keywords: Natural Language Generation, Signal Analysis, Signal Processing, Knowledge-based systems, medical informatics, narrative

*Corresponding author: Albert Gatt, a.gatt@abdn.ac.uk

1. Introduction

Information overload is a pervasive problem in many environments, particularly those in which human decision-making is based on extensive datasets which are collected (semi-)automatically and at regular intervals. For example, contemporary weather forecasting relies heavily on Numerical Weather Prediction (NWP) models, which can generate predictions of several weather parameters (e.g. wind speed), at thousands of different locations at various points during the day [119]. Similarly, gas turbines often have several sensors (for example, to monitor exhaust emissions), which sample data with very high frequency [129]. Such large volumes of data are difficult for humans to digest and interpret. On the other hand, missing important patterns or trends in the data can compromise decision-making, with potentially deleterious consequences.

Currently, the technology of choice for managing such large volumes of information is visualisation. However, there has been a recent surge of interest in *data-to-text* systems, which use Natural Language Generation (NLG) techniques to generate textual summaries of data [101]. This paper discusses a family of knowledge-based data-to-text systems currently being developed within the BabyTalk Project[1], which employ NLG techniques to provide decision support in a Neonatal Intensive Care Unit (NICU)[2]. As patient care standards improve, the demand for continuous monitoring and data collection is on the increase in these units. Therefore, medical staff need to process large quantities of information in order to ensure that clini-

[1] `http://www.csd.abdn.ac.uk/research/babytalk/`
[2] The medical terminology used at various points throughout the paper is defined in the Glossary provided in the Appendix.

AI Communications
ISSN 0921-7126, IOS Press.

cal decisions are maximally beneficial to an infant. The systems we describe aim to reduce this information overload through the use of NLG techniques. Moreover, they target different user groups, namely nurses, doctors and family members or friends. These groups have different information requirements and may also have different levels of expertise.

Although the systems we describe share a number of features with existing data-to-text systems – not least their heavy reliance on domain knowledge – they differ in that they employ a diverse set of techniques to summarise data, including medical signal analysis, knowledge-based reasoning and natural language generation (NLG). Rather than give an exhaustive description of a specific system (for which the reader is referred to [122,96]), our aim in this paper is to discuss the overall vision behind the project and to highlight, through specific examples, a number of theoretical and practical challenges that arise in using NLG for decision support in the medical domain. Two of these are particularly central to the work described here.

The role of NLG in relation to other modalities If the feasibility of NLG technology for decision support in the NICU is to be demonstrated, it must be compared against the current presentation modality of choice, namely information visualisation, various applications of which have been described in a medical context [70]. The main question that arises in this context is therefore *What can* NLG *contribute that is not already provided by a visual presentation?*. We do not intend to argue that NLG should replace visualisation; rather, our intention is to explore the feasibility of using NLG technology in the medical domain, and to attempt to identify some of its contributions. As part of this investigation, we have developed an initial prototype system, BT-45, which was evaluated in an off-ward experiment with clinicians. This experiment compared its utility as a decision support tool to that of a particular form of visualisation which is currently in use in the NICU, as well as to human-authored text. The system and the evaluation have been described in detail elsewhere [122,96]. Here, we summarise the salient points in Section 4, focussing primarily on the evaluation results and their implications for our ongoing work.

Meeting user-specific requirements Staff roles in an ICU are well defined and any new technology has to suit the work flows associated with them [115]. It has been shown that different roles (doctor vs. nurse) and different levels of seniority and experience give rise to differences in staff understanding and use of clinical concepts, suggesting that decision support interfaces need to cater for different user groups [38,36]. The variety of people occupying different roles in the NICU therefore precludes a single, 'one size fits all' solution to the text generation problem and makes it necessary to tailor summaries to the requirements of specific user groups. For example, at the start of a shift, nurses require the sort of information that will help them to plan patient care over the next six- to twelve-hour period, while doctors tend to need information that is directly relevant to decisions about diagnosis and treatment. In addition to medical professionals, another class of stakeholders in the NICU consists of parents, family and friends, whose information needs will depend on how close they are to the patient, and who will tend to prefer a non-technical summary of salient events. Tailoring text generation to these users has consequences for processing at every level, from the stage at which reasoning is carried out to generate abstractions from data, to the stages where content is selected for inclusion in a summary, and rendered as text. The challenges of meeting user requirements are currently being met in the development of three systems, BT-NURSE, BT-FAMILY and BT-CLAN. These are described in Section 5.

The above challenges have some obvious connections to issues that have been topical in the Human-Computer Interaction (HCI) literature for some time, particularly where presentation modalities and user-adaptation are concerned. What distinguishes the work described here from work in HCI is its emphasis on Natural Language Generation techniques, and the underlying hypothesis that language is an ideal modality in which to present information – perhaps in conjunction with other modalities – and adapt it to the requirements of different users. We discuss the motivation for this hypothesis in Section 3.1. Another feature of the present work that distinguishes it from typical HCI approaches is its emphasis on evaluation with actual target users, something that remains a challenge for current HCI methodologies [94]. Our approach to evaluation is made clear in our description of the evaluation experiment for BT-45 (Section 4) and in our current plans for evaluation for the BT-NURSE system (Section 5.1.1).

A related area which is of direct relevance to the present work is that of Clinical Decision Support Systems, which are discussed in Section 3.2. While such systems are designed to impart information, they are different from the systems described here in that (a) they do not tend to involve a significant component

for automatic summary generation; and (b) they tend to be designed to fulfil a recommendation role, pointing out possible courses of action that a user can take given a particular state of affairs. In contrast, the systems described here, while making heavy use of reasoning techniques to perform abstractions from data and to infer relationships between events of medical importance, nevertheless stop short of recommending courses of action, opting instead for a more descriptive summary. As explained in Section 3.2, this is motivated in part by the finding that expert users are resistant to receiving direct recommendations by automated means.

In addition to the above challenges, a project of this nature, which borders on several sub-fields of AI, brings to light several questions of a theoretical and practical nature, whose relevance extends beyond the immediate, domain-specific concerns of the systems being developed. These questions, which we discuss at length in Section 6, fall into two main categories.

Challenges in developing data-to-text systems Systems that bring together information from a variety of sources must be prepared to deal with data in several different formats and integrate them into a single, coherent presentation that will be of some benefit to the end user. The desirability of achieving *coherence* in information presentation imposes requirements at all stages of the NLG architecture. The sheer volume of the data places a heavy burden on the task of selecting the right content and structuring it in a way that maximises the potential benefit of a summary. It also raises the question of how this data should best be presented. Since much of this data is temporal in nature, and relationships such as causal and associative links abound, we are exploring the relevance of work on narrative structure as a way of presenting information in the form of a 'story', which is told with one or more communicative goals in mind. The temporal dimension is therefore crucial; hence, temporal reasoning and reasoning with uncertain information also become relevant, particularly because not all data will be stored and managed with the same degree of accuracy and reliability. Some of this information may even be present in the form of unstructured text, necessitating the use of Information Extraction and/or Natural Language Understanding techniques. We address each of these issues in turn in Section 6.1.

Implications for decision support and information management An important question, which we have already hinted at, relates to what the higher-level goal of a summary should be, that is, whether it should seek to make explicit recommendations, as many current Clinical Decision Support Systems do, or whether it should describe events, perhaps emphasising some over others. Research on the effectiveness of recommendation-based CDSSs has yielded conflicting results [42]. A different approach might be to highlight important events in the data, leaving it up to the user to determine the best course of action. A second, more practical, question relates to whether data-to-text technology, in conjunction with other techniques such as visualisation, holds promise as a way of bringing together information from multiple sites, making them available to a broad, multilingual network of users. We speculate on both of these question in Section 6.2.

The rest of this paper is structured as follows. We will first (Section 2) give a description of the kinds of data that are collected in the NICU and the different ways in which they need to be processed. Here, we also give some examples of medical summaries which are used to motivate our approach. This is followed in Section 3 by a review of related work , focussing in particular on the role of language in information transfer, and the potential role of NLG as a decision support technology, in comparison to visualisation and expert systems. We then address the challenges outlined above, first describing our initial prototype system (Section 4), and then giving an overview of work in progress on the systems that are under development for the main user groups identified above (Section 5). Finally, in Section 6, we address the broader challenges and their implications against the background of our ongoing work. Section 7 concludes with some remarks about future directions.

2. Data in the NICU and the BabyTalk vision

A patient in the Neonatal Intensive Care Unit is usually a premature infant with health complications, who requires life support, continuous monitoring and treatment for a period ranging from a few weeks to several months. Contemporary NICUs collect large quantities of data about these infants. In the drive to enhance patient safety and improve decision making through accurate record-keeping, much of this data is being stored electronically and is accessed routinely by doctors and nurses in the course of a shift. The NICU data that is the focus of the BabyTalk project comes from the Neonatal Unit at the Edinburgh Royal Infirmary, one of the project partners, and is collected through an electronic

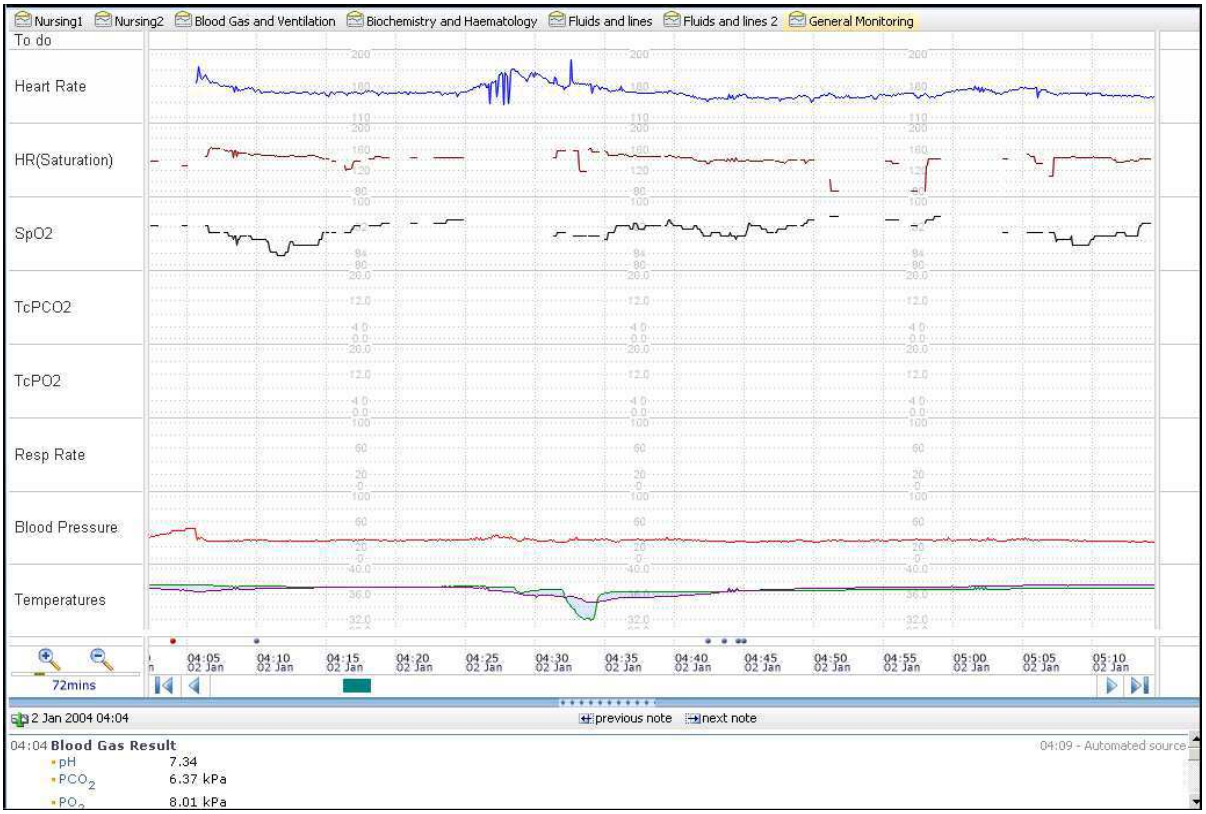

Fig. 1. Time series display of physiological information. The bottom panel of the interface displays database entries at the time being displayed.

medical record (EMR) system called Badger 3^{TM}, built by Clevermed Ltd [22]. The system is accessed by doctors and nurses at the cotside. The data collected and displayed by the system falls into a number of categories, described below.

Continuously sampled physiological data Values from different physiological parameters (referred to as *channels*) are sampled automatically via sensors, typically at a rate of 1Hz. Channels include Heart Rate (HR), mean blood pressure (Mean BP) and Oxygen Saturation (SPO2). Given the sampling frequency, the data from a particular channel constitute a high-density time series; the total amount of physiological data collected from a single patient over a single 12-hour shift can run into several megabytes. Figure 1 shows an example of this information, as displayed by the Badger system. The figure displays time series plots for 9 different physiological channels (the panel labelled *temperatures* actually plots both core and peripheral temperature), only six of which are actually being recorded.

Numeric data from lab results and observations Another kind of numeric data in our domain comes from

Blood Gas (04:04)	
PH	7.34
CO2	6.37
PO2	8.02
BE	-0.3
Haemoglobin	18.7
...	...
Glucose	3.8
Lactate	1.7

Fig. 2. Example of sporadically sampled numeric data

laboratory results and observations that are entered in the database sporadically. These data can also be viewed as a time series, but in contrast to the continuously sampled data, they are not as dense, and may be sampled at irregular intervals. As an example, Figure 2 gives the details of the results of a blood gas analysis. This is a laboratory test used to measure various parameters in the blood and is taken at regular intervals, with results entered in the database. These data can also be displayed with the time series plots as shown in the bottom panel of Figure 1; in this instance, a blood gas result is noted to have been received at 04:04. This

class of data also includes a series of hourly observations made about the patient, recording parameters such as respiratory rate and other values related to a baby's respiratory support.

Ventilation information

Ventilation	CMV
Phototherapy	single
Intravenous fluids	amino acids
Adequate urine volume	yes
Stools	meconium
Drugs	Benzypenicillin
	Gentamycin
	Nystatin cream
	Nystatin suspension

Fig. 3. Example of sporadically recorded symbolic data

Symbolic data Electronic forms are used to store information about a variety of events during the course of a shift. These include nursing actions, medical diagnoses and treatments, and information related to a baby's parents and guardians. Though sparser than the numeric and physiological data, the variety of types of information stored nevertheless make the amount collected over a single shift significant. As an example, Figure 3 displays various items of information, including the baby's current respiratory support, whether or not the baby is undergoing phototherapy, the medication s/he is on, and details about feeding and fluids.

Free text data In addition to structured input, doctors and nurses can also enter observations and comments in the form of free (unstructured) text. These notes often include observations about a patient's state, such as whether s/he is stable or distressed. Occasionally, the free text notes serve to give some justification or explanation for events entered through the structured forms. An example of a free text note for this patient is shown below. This note makes observations related to the patient's respiration and circulation.

> Had desaturation down to 65's, FIO2 increased up to 29%, colour slightly dusky, takes time to recover.

2.1. Record-keeping versus integrating information

Record-keeping systems such as the one described above make data storage and retrieval highly efficient. However, the way in which information can be accessed has certain limitations.

One limitation, already noted in Section 1, is related to the modality of presentation. The visualisation of physiological data does not highlight important patterns or trends in the time series, placing the burden of discovery on the viewer. This means that it is up to the nurse or doctor to spot important events such as sudden drops in heart rate, or fluctuations in temperature. While this is relatively easy for an experienced person, it is a skill that tends to be acquired over time and relatively inexperienced staff may be more likely to miss important information. Moreover, even experienced doctors and nurses are likely to make occasional errors [3].

A second, perhaps more important, limitation is that information is available in a piecemeal fashion, that is, the user can look up the information under specific headings, using different forms and displays, but there is no single location which brings together all potentially related data items, telling a coherent story which highlights clinically important events. A good example of this involves nurse shift summaries, which are produced at the end of every twelve hour shift. The current patient information system partially overcomes this limitation in that it automatically generates a shift summary report which brings together several data items, as well as additional notes entered manually by a nurse. Information is presented under specific headings, broadly corresponding to different physiological systems and functions, such as *respiration*, *circulation*, and *fluids and feeds*. An example summary, covering the same period from which the previous examples were drawn, is shown in Figure 4(a). This shows the parts of the summary that are relevant to a patient's respiration, though, as shown in the Figure, some information appears under the headings *Circulatory*, *Notes* and *Other*. The information under these headings is displayed on different parts of the screen, separated by several other pieces of information pertaining to other topics. Moreover, the selection of the data items collected is static - it does not depend on the current clinical context. So, for example, what is included or omitted does not depend on its current importance, or on whether it has any bearing on treatment decisions. Also, there is no way to highlight trends and patterns in time series unless these happen to be included in a free text note and then included in the manually entered section of the shift summary.

The summary format in Figure 4(a) is fairly standard, and has the benefit of organising information in a way familiar to a nurse, while permitting easy lookup. However, presenting information in such a compart-

Respiratory
Respiratory support: CMV
Inspired oxygen: 26.00 % or lpm
Oxygen range % From:21 To: 26
Oxygen saturation range From:88 To: 94
Respiratory notes: On CMV pre-16/4, FIO2-26%, BR-20bpm.

Notes
Quite settled night. Active, pink. Remains on CMV mode overnight. [...] BLd.gas done and acceptable. Had 1x desaturations down to 60's, colour slight dusky needing increased FIO2 and pressure increased to 16/4.

Cirulatory
Had desaturation down to 65's, FIO2 increased up to 29%, colour slightly dusky, takes time to recover. Informed the doctor and vent settings adjusted, pressure increased to 16/4. Suctioning done obtained 1 mucoid dirty secretions from ETT and orally.

Other
22:40 SBR in bld.gas 98, started on single phototheraphy.
23:49 Bld. gas done/seen, ventilator setting adjusted pre-14/4.

(a) Excerpts from a nurse shift summary displayed by the system in use at the NICU, which brings together free text notes and other data on a single screen. Parts headed in boldface correspond to different sections which appear on different parts of the display as separate topics.

Breathing
Current management: Currently ventilated on CMV, rate 20, pressures 16/4, iTime 0.3 seconds, in 26% oxygen. ETT size 2.5 is 6 cm at the lips.

Current assessment: Respiratory effort reasonably good, his total resp rate being 40-50 breaths/minute while the ventilator rate is 20. Tidal volumes are 2 - 2.8ml on the current settings. CO2 on the last gas (04:00) was 6.37kPa. ETT and oral suction yielded small amount of mucoid but stained secretions.

Events during the shift: Baseline SpO2 had drifted down from 95% to 88% accompanied by increasing SpO2 variability, HR stable. After blood gas at 23:00 ventilation pressure reduced to 14/4. CO2 was 4.1 and tidal volumes were 3.8 - 4ml at that time. After a desaturation 3 hours later down to 65% pressures were put back to 16/4 He has had an oxygen requirement of 26% since this episode.

Potential problems: Small ETT could become blocked or dislodged - ongoing assessment of need for suction; ensure ETT is secure. Risk of chest infection related to being ventilated; also due to extreme prematurity and PROM he is at risk of ureaplasma infection - daily ETT secretions samples should be sent for C&S and ureaplasma.

(b) Excerpt from a retrospective summary written by a senior nurse for the same period. The excerpt corresponds to a single section pertaining to a single topic, ventilation, with subheadings as marked by the author.

Fig. 4. An actual (a) and retrospectively written (b) nursing shift summary.

mentalised fashion does not always allow the relationship between different kinds of data to be highlighted. This is already evident from the fact that information related to respiration is sometimes found under alternative headings (because it is also related to other physiological systems). Other differences are thrown up more clearly through a contrast between this summary and that shown in Figure 4(b). This is an alternative shift summary report for the same period, written retrospectively by a senior neonatal nurse. It forms part of a small corpus of such summaries collected as development data for one of the BabyTalk systems, BT-NURSE (Section 5.1). In addition to basic information about the patient's current respiratory support, both summaries mention, among others, the following events:

- blood gas results, including the result displayed in Figure 2;
- fluctuations in Oxygen Saturation, including mention of desaturations;
- changes to ventilator settings, such as FIO2 and ventilator airway pressure;
- suction, with details of secretions.

However, this information is presented somewhat differently. From the point of view of presentation, the summary in Figure 4(b) groups all this information under the single heading *Breathing*. Moreover, it begins with an overview of the patient's current state, comprising an assessment of respiratory effort based on the ventilator parameters. This section also mentions ETT Tube size, which is later flagged up as a possible area of concern under the sub-heading *Potential Problems*. In addition, the CO2 value from the last blood gas is mentioned, because this is an important indicator of acidity levels in the blood, as well as the extent to which the patient is making an effort to breathe independently. Another important difference between the two summaries is that the retrospective one highlights several causal links between events, such as the fact that the ventilator settings were adjusted as a result of a blood gas result, and again due to desaturations. The

same information is mentioned in the original summary under two different headings (*Notes* and *Circulatory*), which the system displays in different parts of the screen.

There are a number of differences between these two ways of presenting information. First, most of the information presented under the four different headings in Figure 4(a) is brought together in the retrospective summary under the single heading *Breathing*. Second, the summary in Figure 4(b) makes reference to trends in the physiological data to a much greater extent than does the one in Figure 4(a), particularly in the section entitled *Events during the shift*. The trend data can of course be viewed independently on the display shown in Figure 1, but the relationship to other actions and events is not always made explicit in the original summary of Figure 4(a), save for some mentions of desaturations under the heading *Circulatory*. Similarly, the summary in Figure 4(b) gives a more detailed account of the baby's current state, with reasons for the assessment, also highlighting potential problems due to care actions taken in the course of the shift. The second summary also links the baby's current state to the medical history since birth, noting that the current ventilator settings can give rise to infection because of the baby's prematurity.

2.2. Potential benefits of automatic summarisation

The differences between these two approaches to data summarisation seems due in part to a difference in their authors' motivations. The original summary shown in Figure 4(a) is primarily intended to impart information about the main events of the last twelve hours, in a format that is based on principles of organisation which have a basis in physiology, and which is intended to facilitate easy lookup on the part of a nurse beginning a new shift. However, an indirect result of this structure is that relationships between different parts may be missed. It is worth emphasising that such a summary is normally accompanied by a verbal shift handover, where additional information is supplied.

In addition to imparting information, the alternative summary in Figure 4(b) also stems from an intention to present a coherent *narrative* about what the patient has undergone in the last 12 hours. What distinguishes narrative discourse from mere representations of facts is that it involves a construal of events which emphasises some as being more important, thus communicating a point that goes beyond the facts themselves [87]. To achieve this, a text must satisfy several requirements, many of which have been identified in work on narrative discourse both within the tradition of discourse studies – especially the work of Labov [73,74] – and within psycholinguistics [133,132]. The first of these concerns the temporal dimension: the summary enables the reader to reconstruct events in time, using such mechanisms as tense and aspect features and adverbial modifiers which either situate events in time in an absolute sense (*at 23:00*), or relative to other events (*3 hours later*, after blood gas). More crucially, the text supports the inference of causal relations, a central aspect of narrative [47]. In our example, this is evident in such formulations as

> *After blood gas [. . .] ventilation pressure reduced to 14/4. CO2 was 4.1 . . . .*

where the text supports the inference that the CO2 levels found in the blood gas warranted the change in ventilation pressure. Finally, the reader is also able to infer the communicative goal or purpose that informs the choice of content in the summary. The fact that only a subset of the events during the previous 12 hours is mentioned implies that the selected subset is relevant to the overall communicative goal. Thus, considerable emphasis is placed on salient patterns in the physiological signals (desaturations, changes in SpO2) and changes in the oxygen requirement of the baby. This may cause someone who is starting on a new shift to pay particular attention to the baby's oxygen requirements and ventilator pressure. By way of conclusion, the section on potential problems identifies areas of concern which are directly linked to the narrative.

Since NICU data are stored in several different formats and the volume is considerable, sifting through all the available information to construct such a narrative summary is extremely time-consuming. For example, the summary of which an excerpt is shown in Figure 4(b) took an expert several hours to write. Therefore, such an exercise is not currently feasible in real-time patient care. This is where NLG techniques can play a crucial role. On the other hand, the automatic generation of such texts raises many linguistic challenges, not least the ability to structure the text coherently, allowing the reader to keep track of time shifts and cross-references to entities mentioned at several point in the discourse [133,131].

Our aims in this project are (a) to build systems that can quickly and effectively present a narrative summary; (ii) to evaluate these systems with their target users, testing their effectiveness in clinical decision-

making compared to other ways of presenting information. Before turning to the details of how this vision is being achieved, we first discuss some of the precedents for the work presented here.

3. Related work

This section gives an overview of some related work, focussing on the fields of Information Visualisation, computerised Clinical Decision Support Systems (CDSS), and Natural Language Generation (NLG). Our focus will be on the evidence for the utility of these tools in helping decision-making.

3.1. Information Visualisation and decision support

Information Visualisation techniques aim to provide users with effective means of presenting, exploring and interacting with large datasets, reducing the complexity of examining and understanding such data [19]. An influential taxonomy of visualisation techniques classifies them orthogonally by task and data type [109]; this classification includes time-oriented data (such as the patient data in Figure 1) as one of the types.

Visualisation techniques for time series data have focussed on challenges such as presenting high-density data with limited resources (e.g. limited screen resolution) [2,84], dealing with unevenly sampled data [6], and allowing interactive search and zoom functionality [15]. Within a medical context, a survey by Kosara and Miksch [70] has identified various methods of representing quantitative data, such as charts and graphical patient record summaries, which give a complete overview of several patient parameters, with the display resolution reflecting the recency of events [97]. Interestingly, none of the methods surveyed is judged to satisfy the full set of requirements that the authors identify for such systems. These include an ability to combine multiple values on a single display, while permitting the user to identify salient patterns and intervals. In addition, few visualisation techniques for time-oriented data can handle a combination of quantitative and discrete information[3].

There is some psychological evidence in support of the effectiveness of visualisation. Among other factors, visualisation facilitates visual chunking [128], and also reduces memory load [110]. However evaluations of visualisation techniques have largely taken place in the laboratory, rather than in real settings, focussing on usability-related issues [94]. How effective novel visualisation techniques have been in medical decision-making is therefore harder to assess.

On the other hand, there has some research on more traditional ways of visualising data. Elting *et al* [34,35] compared the effectiveness of pie charts, tables, icon displays and text in viewing and interpreting the results of clinical trials. The outcomes showed that physicians performed better with icon displays, though these were not their preferred modalities. However, these studies focussed on relatively low-density data, consisting of mortality rates; moreover, it is not clear what kind of text was actually used in the experiments [34].

There have been some recent challenges to the effectiveness of visualisation techniques of high-density clinical time-series data, particularly in the NICU context. Cunningham *et al* [24] reported a study showing that clinical outcomes were not improved by an implementation of a trend monitoring system in the NICU. In this connection, McIntosh *et al* [79] analysed recordings of a number of doctors and nurses at different levels of seniority and experience, as they described their observations of patterns and trends in time-series visualisations of medical data. The analysis suggested that expertise plays a crucial role in the extent to which clinicians identify significant patterns and observe possible links (for example, causal links or correlations) between different channels. However, even senior doctors tended to miss patterns in the multichannel data which should give early warnings of impending problems for a patient. These findings were echoed by Alberdi *et al* [3], who found that success at detecting crucial trends and patterns depended on experience with the monitoring device. Second, senior doctors were far more likely than juniors to detect artifacts (noise) in the trend data, something which many visualisation techniques are not equipped to detect. A third issue concerns the way in which clinicians bring to bear medical knowledge on their interpretation of trends and the inferences they make from them. This is a task that senior clinicians are better equipped to carry out, suggesting that automatic methods of identifying links in the data would be particularly useful for junior staff.

Many of the problems identified by these studies in relation to trend data monitoring are precisely those that served as the motivation for the work reported here. Automatic summarisation has the potential to highlight patterns and trends (and describe them as important or unimportant), linking them to non-

[3]The survey authors refer to these data as incidents and symptoms. They are distinguished from time series data in that they do not involve regularly sampled numerical values.

quantitative (discrete) events; moreover, if preceded by a signal analysis stage which includes artifact detection, a summary can avoid emphasising patterns which are of lesser clinical importance. Thus, one would expect a natural language summary to be able to overcome some of the limitations of the sorts of visualisation techniques currently in use in the NICU, such as that shown in Figure 1. This was in part confirmed by a more recent experiment by Law *et al* [75], which explicitly compared clinical decision making based on NICU data presented either in the form of a textual summary (written by expert neonatologists) or using graphical displays. Participants were asked to select the appropriate courses of action to take in relation to a patient, given the data. The results showed that better decisions were taken when information was presented textually, rather than graphically. Interestingly, clinicians themselves stated a definite preference for the graphical modality. The discrepancy between preference and actual performance echoes findings by Elting *et al* [34].

3.2. Clinical Decision Support Systems

Perhaps the most obvious way to use artificial intelligence techniques for decision-support is to explicitly give advice to users. In medicine, AI researchers have been working on Clinical Decision Support Systems (CDSSs) since the 1970s, when the MYCIN system (which diagnosed bacterial infections and recommended appropriate treatments) was shown to give better advice than many doctors [14]. Many CDSSs have been developed since this time, ranging from expert rule systems to model based systems; however, they have not been widely integrated with routine workflow as decision-support aids. This is in part due to the fact that there is little evidence that such systems actually enhance patient outcome. CDSSs can be evaluated either by comparing decision quality against a gold standard (e.g., are doctors assisted by a system more likely to make the 'right' decision as defined by a gold standard), or by measuring their impact on patient outcome (e.g., mortality). Garg *et al* [42] reviewed evaluations of CDSSs and concluded that there is good evidence that CDSSs *can* enhance decision quality from a gold-standard perspective, but that the evidence that they actually improve patient outcome is much weaker. Garg *et al* also observed that evaluation studies conducted by the system developers were three times more likely to observe improvements in decision-making against a gold standard than evaluation studies conducted by other people. Tan *et al* [116] reviewed CDSSs specifically for neonatal care, and concluded that there was currently no evidence that decision-support systems for neonatal care improved patient outcomes (largely because very few studies of neonatal decision-support systems had measured these). This situation mirrors that discussed by Plaisant [94] in relation to visualisation, whereby novel techniques tend not to be evaluated extensively in real settings.

One possible explanation of the limited use of CDSSs in practice could be that many developers have not directly addressed the question of how users react to being directly advised by an automated system. In addition, there have been concerns about the legal implications of deploying such systems. For example, legal responsibility for medical decisions remains with doctors, and many doctors are reluctant to take responsibility for decisions based on recommendations made by a piece of software which they do not understand or which does not explain the rationale for a conclusion, or the level of confidence with which it is reached. In fact, many systems have been abandoned because of the ratio of false alarms they generate [18] which tends to remove the confidence clinicians may have in the machine. Another point is that there are only a few studies which consider usability assessment as part of the system evaluation [77], whereas a decision support system has the ultimate goal of communicating with humans.

Tehrani and Roum [117] present a very recent (2007) review of decision-support for ventilation decisions, which are very important in BabyTalk's NICU domain [54]. They review 21 systems in this domain, developed between 1985 and 2007, which used a variety of AI techniques (including rule-based reasoning, model-based reasoning, fuzzy reasoning, and temporal abstraction). They summarise several promising evaluations of such systems, but it is difficult to interpret these because they do not screen the evaluations according to their methodological rigour. However, they make the interesting observation that while expert system ideas have been incorporated into a number of commercial 'closed-loop' systems (that is, control systems which automatically adjust ventilator parameters without human input), decision-support systems which advise human doctors have not been widely used in operational commercial systems. This trend has also been confirmed in a recent development in the sonography domain. An evaluation of the SONOCONSULT system [98], which provides both a diagnostic aid and data management support, has shown that while clinicians

accept the data management support function, the diagnostic aid is not used, although the clinicians recognized its correctness.

The above surveys present a contradictory picture. On the one hand, AI medical reasoning techniques do seem to work at least to some degree, but on the other hand CDSSs which give medical advice have not proved successful. In a sense, the BabyTalk project can be considered as an attempt to support medical decision in a different way, by summarising data as text instead of giving direct advice. Internally, Babytalk uses many of the same AI reasoning techniques as the systems described by Tehrani and Raum [117] (including signal analysis, artifact detection, and production rules); but it uses these techniques to help doctors understand and interpret data, without telling clinicians what to do or giving high level diagnoses. Our hypothesis is that this use of medical reasoning will be more successful than advice giving, in part because it is more robust in the face of data problems (noise, missing data) and incomplete knowledge bases, and because it is more acceptable to doctors (who prefer to retain control over what to do). On the other hand, we are also currently facing the question of whether the system should include at least *some* advice or recommendation in the generated output, a topic to which we return in Section 6.2.2.

3.3. Natural Language Generation and data-to-text technology

Natural Language Generation (NLG) systems, which produce text or speech from non-linguistic input [102], have been deployed in a variety of domains. Recent examples include interactive museum guides [88,114] and generators that render scripted dialogues between avatars [121]. One class of systems which is particularly relevant to the present work is that of *data-to-text* applications [101]. Such systems produce summaries of numeric and symbolic data which require one or more pre-processing stages in order to extract the information required for generation, thereby necessitating a data analysis and/or interpretation stage prior to the stages typically associated with NLG proper [102]. This is one of the main features distinguishing data-to-text applications from other kinds of NLG systems, whose input tends to consist of information that is already structured in some way, though several stages will of course intervene between the input and the final text.

The domains in which data-to-text systems have been deployed vary greatly, ranging from summarisation of statistical data [61,37], to stock market trends [72,25] and environmental data [13]. More recently, these techniques have also been used as aids for users with special communication needs, by generating summaries of users' daily activities which are collected from sensors [106]. Weather reporting has also featured strongly in the history of data-to-text systems and has been among the success stories in NLG, starting in the mid-nineties [46,23]. More recent weather forecasting systems have been extensively evaluated. For example, SumTime [112,105] was shown to generate forecasts which were preferred by human readers over those produced by professional forecasters. This was arguably the first such evaluation, and has been followed up by others in a related vein [11,10].

Many data-to-text systems tend to work with relatively small datasets. Exceptions to this trend include RoadSafe [119], which works with very large sets of geo-referenced meteorological data, and the system described by Yu *et al* [129] to summarise patterns in gas turbine sensor data. However, both systems assume fairly homogeneous input data, largely consisting of a single datatype. In this respect, the systems discussed in this paper, including the BT-45 prototype described in Section 4, are novel insofar as they handle very large datasets consisting of different datatypes (see Section 2), generating multi-paragraph texts for which non-trivial solutions are required at every stage of generation.

In the medical context, there has been a significant focus on summarisation from textual resources [1], but data-to-text technology has had a more restricted application [57] and existing systems have largely focussed on discrete data (e.g. [64,56,51]). These systems have either generated patient reports for medical practitioners, or produced text that was targetted at the patients themselves [104,17]. Evaluation of these systems has yielded mixed results. For example, the STOP system [104], which generated smoking cessation letters, was evaluated in a clinical trial which suggested that the generated letters were not effective in motivating readers to stop smoking.

In summary, data-to-text technology has been shown to be viable in some domains, particularly weather forecasting, but has yet to be shown to be a feasible complement to existing decision support technologies in medical contexts. Moreover, with the exception of STOP, existing systems have tended to be evaluated by eliciting judgements from human users, rather than by

assessing task performance based on their output. The system described in the following section, BT-45, is probably the first system of its size and scope to generate summaries automatically from a diversity of types of data. Its evaluation also focussed on task performance.

4. BT-45: A prototype and its evaluation

In this section, we describe BT-45, a system that generates summaries of patient data spanning approximately 45 minutes. This system was intended as a preliminary test of the feasibility of NLG technology for neonatal decision support. An additional motivation came from the experimental results reported by Law *et al.* [75] (see Section 3.1), whose results, showing that textual summaries result in better decision-making than existing techniques for graphical data presentation, also used scenarios of approximately 45 minutes of patient data in the form of physiological signals and sporadically recorded events. A replication of these results, comparing both human-authored and automatically generated summaries to the graphical presentations, would therefore serve both to evaluate an initial prototype as proof of concept, and also to test the robustness of the findings from Law *et al.*'s study. As in the Law *et al.* study, the database of discrete events used for the development of BT-45 contained, in addition to data routinely logged by medical staff, annotations of clinical events in the NICU recorded by a research nurse over the course of a few months as part of the NEONATE project [60].

From the point of view of the overall aims of the BabyTalk Project, the design of BT-45 also led to the development of the core components and algorithms that are currently being ported to the BabyTalk systems described in the next section. The same is true of the domain knowledge that subtends the systems currently under development. Knowledge acquisition in BabyTalk has largely been based on extensive interviews with clinical experts. Starting with BT-45, these have served as the starting point for the development of an ontology of NICU concepts, described in Section 4.1.1, as well as the expert rules used by the reasoning component (Section 4.1.3). These have since been extended further to meet the requirements of the new systems, particularly BT-NURSE (Section 5.1).

A complete description of BT-45 has been given in Portet *et al* [96]; the evaluation experiment is also discussed in detail in van der Meulen *et al* [122]. Here, we give an overview of the system architecture, focussing on the main challenges in building such a system and on the lessons learned from the evaluation.

Fig. 5. Human authored summary of a 45-minute data period

> You saw the baby between 16:40 and 17:25.
>
> Initially the HR baseline is 140-160; pO2 is 8-10; oxygen saturation = 92%, T1 and T2 are 36.9 and 36.6C. At around 16.45 ET suction is performed; there is a drop in oxygen saturation to 50% and pO2 to 3.3 and a rise in pCO2 to around 9. The FiO2 is increased from 61 to 100%. By 16.51 the HR is at 155, the pO2 is 6.7 and the oxygen saturation is 88% and the pCO2 is 9.2. There is an upward spike in the pO2 to 16.9 and a corresponding downward one in pCO2 to 3.1; the oxygen saturation has fallen to 78%. T1 is now reading 36.9C and T2 35.7C. At 16.57 the ventilator rate is increased to 30.
>
> Baby is given Neopuff ventilation. The oxygenation continues to decrease: pO2 = 0.2 and oxygen saturation = 20% at 17:00 and the HR falls to 60. The pCO2 continues to rise to 10.1. The baby is pale and unresponsive. ET suction is given, baby is turned and at 17:02 the ETT is removed; the baby is again given Neopuff ventilation. Baby is re-positioned and the NGT aspirated. By 17:08 the baby is reintubated; the oxygen saturation has increased to the 80s and the HR has risen to 176 the pO2 = 0.1 and pCO2 = 0.2, T1 is 32.7C and T2 34.7C.
>
> At 17:15 the FiO2 is reduced to 33% and the rate put back to 15.
>
> At 17:24 the oxygen saturation falls to 65 and the FiO2 is increased to 56%.
>
> At 17:25 the HR is 165, the oxygen saturation is 100%, T1 is 35.7 and T2 is 34.5C.

Figure 5 displays an example summary used in the BT-45 evaluation. This is a consensus summary written by a senior neonatologist and a senior neonatal nurse. Figure 6 is the corresponding summary generated by the BT-45 system from the same data. The data itself consists of physiological signals, numeric results, and discrete events only; that is, BT-45 does not incorporate free text data in the sense discussed in Section 2. Further, the summaries are primarily *descriptive*. Data interpretation (both in the human and the automatic summaries) is confined to (a) interpreting patterns in the physiological signals, as in the identification of bradycardias and desaturations; and (b) linking events, particularly those which are causally connected. Thus, there is no attempt to diagnose problems or recommend courses of action but only to present information in a way that would facilitate the process of deciding on these factors by a doctor or a nurse.

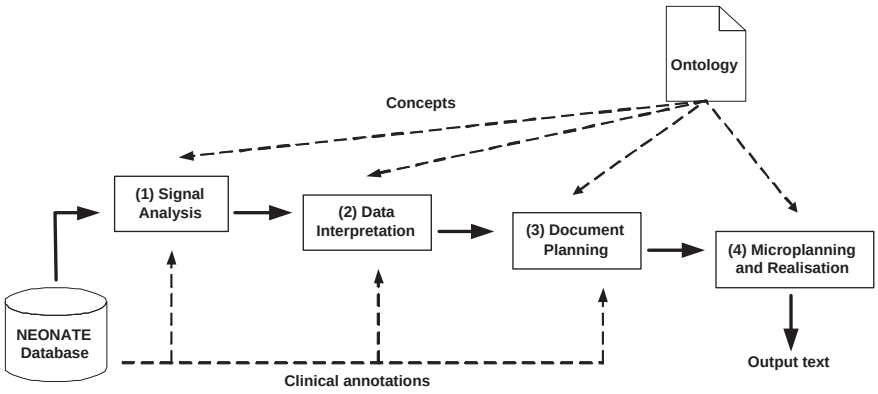

Fig. 7. Architecture of the BT-45 system

Fig. 6. Automatically generated summary by BT-45

> You saw the baby between 16:40 and 17:25. Heart Rate (HR) = 155. Core Temperature (T1) = 36.9. Peripheral Temperature (T2) = 36.6. Transcutaneous Oxygen (TcPO2) = 9.0. Transcutaneous CO2 (TcPCO2) = 7.4. Oxygen Saturation (SaO2) = 94.
>
> Over the next 24 minutes there were a number of successive desaturations down to 0. Fraction of Inspired Oxygen (FIO2) was raised to 100%. There were 3 successive bradycardias down to 69. Neopuff ventilation was given to the baby a number of times. The baby was re-intubated successfully. The baby was resuscitated. The baby had bruised skin.
>
> Blood gas results received at 16:45 showed that PH = 7.3, PO2 = 5, PCO2 = 6.9 and BE = -0.7.
>
> At 17:15 FIO2 was lowered to 33%. TcPO2 had rapidly decreased to 8.8. Previously T1 had rapidly increased to 35.0.

4.1. System overview

BT-45 follows the architecture for data-to-text systems described by Reiter [101]. The system architecture, shown in Figure 7, consists of four main modules all of which are driven by information stored in the knowledge base of the NICU Ontology.

The different representations handled by these modules and the role of the ontology in providing a common vocabulary can be illustrated with reference to Figure 8. This figure corresponds to the clause in Figure 6 that says *There were 3 successive bradycardias down to 69.*, for which the relevant input is the Heart Rate (HR) physiological signal. Signal analysis identifies the fluctuations in the signal indicated. During data interpretation, these drops in HR are identified as bradycardias. Additionally, they are assigned an importance value, based in part on the lowest value reached by HR during the event, and grouped into a sequence due to their temporal proximity. The next stage is Document Planning, which selects the content to mention, and structures them into an initial document plan in the form of a tree whose edges relate different events to eachother. In the current example, only those elements of the sequence that are sufficiently important are selected. The corresponding part of the document plan consists of a temporal sequence (TSEQUENCE) node, which is linked to its elements via an INCLUDES relation. The microplanning component has the task of mapping the selected events in the document plan to semantic representations. For the current example, the microplanner selects *be* as the main predicate; the representation also includes the main arguments (here, the THEME and VALUE; see Section 4.1.5 below)). The semantic representation is mapped to a syntactic structure, and finally linearised as text by the realiser. In what follows, we describe each of these processing stages in more detail. We begin with an overview of the NICU ontology.

4.1.1. NICU Ontology

One of the novel features of this architecture is its use of an ontology which captures much of the domain knowledge required by the reasoning and data interpretation module (in the form of concepts and properties, and relations between concepts), and some of the linguistic knowledge needed to map this to a linguistic representation. Additional knowledge is encoded in the reasoning rules used by data interpretation. Such a unified repository of knowledge ensures that all the components of the system can communicate effectively, in spite of using very different data structures internally. Moreover, as shown in Figure 7, the ontology is programmatically integrated into the system. It was implemented in Protégé-Frames 2000 [85], which provides

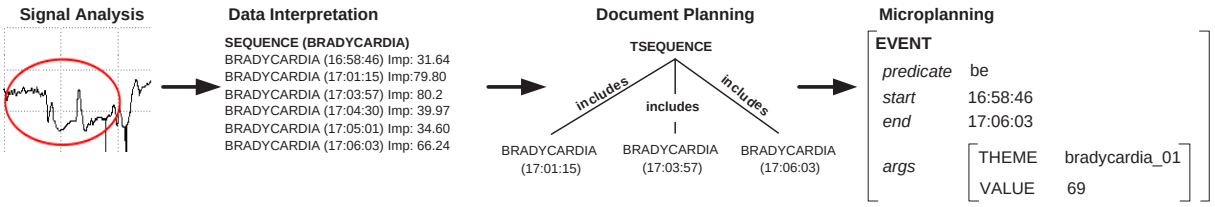

Fig. 8. Mappings between different representations in BT-45

a Java API and can be integrated with the JESS production rule system [40].

The ontology was developed from the ground up, based on a NICU lexicon which specified the words used by nurses and doctors to talk about the NICU domain [60], with additional concepts required by the new application. Its structure was partly informed by that of the Unified Medical Language System (UMLS) [58].

The primary distinction within the ontology is that between EVENT and ENTITY. The former subsumes concepts related to medical interventions (such as DRUG ADMINISTRATION), observations (such as the baby's having poorly perfused skin), and patterns in physiological signals, including TREND and SPIKE. ENTITY subsumes concepts corresponding to a variety of 'non-temporal' objects, such as human actors (e.g. BABY and NURSE), instruments (VENTILATOR), and medication (MORPHINE). All events have slots (features) which specify various properties, such as the peak value of a TREND. In addition, since the ontology also supports language processing, events have properties corresponding to their participants; these are mapped at the microplanning stage to thematic roles (such as AGENT and RECIPIENT), which are realised as the noun phrases in the clause corresponding to an event. This is what enables the microplanner to identify the arguments of the main predicate in the event representation of Figure 8. As another example, a DRUG ADMINISTRATION event has slots corresponding to the patient, who is receiving the drug (and is mapped to a RECIPIENT role), the drug itself (linguistically, the THEME) and possibly the person administering the drug (the AGENT).

4.1.2. Signal Analysis

Signal Analysis processes the physiological data to identify and extract events which are not represented as discrete entries in the NICU database. It proceeds in two main steps: *Artifact Detection* and *Event Identification*.

Artifacts are segments of the input signal that do not correspond to actual values. They may arise for a number of reasons, such as a sensor becoming disconnected due to a baby's being handled. Artifact Detection identifies and removes these based on three techniques: (i)*range checking* identifies values that are not physiologically plausible; (ii) *autoregressive modelling* flags values outside a dynamically updated acceptance interval, while (iii) *correlation checking* uses domain knowledge to relate artifacts found in different channels. For example, TcPO2 and TcPCO2 signals are obtained from a single probe, so that noise in one signal generally implies noise in the other.

Event Identification identifies sequences of consecutive data points in the signals which are of medical significance. These include patterns corresponding to medically known events (e.g. the bradycardias in Figure 8), which are identified using a threshold method. Short-term patterns, which occur over a period of 30 seconds or less, are identified using the rapid-change detection algorithm described by Yu *et al.* [129], whereby adjacent fluctuations in a signal are classified and merged. These are further classified in terms of their direction (rising, falling, steady or varying) and magnitude and/or speed. Such a pattern is exemplified in the final paragraph of the text in Figure 6, where TcPO2 is described as having *rapidly decreased to 8.8*. In addition to short-term patterns, long-term trends (i.e. trends on a time-scale of minutes rather than seconds), are also identified using bottom-up segmentation [66].

Following Event Identification, each event is assigned an importance value. In the case of events extracted from signals, this is carried out using a combination of expert system rules and range value modelling. Thus, the bradycardias in the sequence in Figure 8 have different importance based on their lowest values. Discrete events are assigned importance values based on domain knowledge encoded in JESS rules, and collected through consultation with Neonatology experts.

4.1.3. Data Interpretation

All the processing carried out by the Data Interpretation phase is rule-based and involves both *Temporal*

Abstraction and *Linking*. Temporal Abstraction recursively groups together any two events of a particular ontology class which share certain features and occur within a specific temporal window, abstracting such sequences into higher-level events. This process gives rise to the sequence of bradycardias in Figure 8. Another set of rules are used to create a higher-order event from a set of events of different ontological types. For example, the discrete data may contain multiple INTUBATION and EXTUBATION events within a short period. These are abstracted into a single INTUBATION event, possibly followed by a RE-INTUBATION (if the baby has been extubated and then intubated again), or an EXTUBATION.

Linking is carried out using production rules which fire in response to certain events and conditions. For example, a sudden drop in heart rate (a BRADYCARDIA) may have been caused by the administration of a dose of morphine which happened a short while before. In this case, the rule in question would associate these two events using a CAUSES link. Other links include ASSOCIATE, for events which are correlated, and INCLUDES for part-of relations. These links are also used by the Document Planning module to relate parts of the document to eachother.

4.1.4. Document Planning

Document Planning performs the task of *Content Selection* and *Document Structuring*, outputting a Document Plan. Document Plans are labelled trees whose nodes are messages and whose edges are labelled with the discourse relations between the messages. Thus, the TSEQUENCE node in Figure 8 is linked to three daughters via an INCLUDES relation; other relations include CAUSES and REASON (roughly corresponding to the case where one event is the motivation, rather than the direct cause, of another).

Processing in this module is entirely rule-based and controlled by various parameters specifying the maximum length of a document and the minimum importance an event must have in order to be mentioned. Content selection chooses from among the events output by the Signal Analysis and Data Interpretation stages, using an algorithm based on that of Hallett *et al.* [49]. Each selected event is considered a single 'message' or content unit, and forms a node in the resulting Document Plan. The algorithm identifies *key events*, which are typically those whose importance exceeds a preset threshold. Each of these heads a paragraph, and paragraphs in a document are ordered by the time of occurrence of the key events. Other events are included in a paragraph if they are explicitly linked to the key event or occur at approximately the same time.

4.1.5. Microplanning and Realisation

Microplanning is usually defined as the process of planning the linguistic (semantic) content of messages [102]. It is distinguished from *Realisation* in that the latter is typically concerned with mapping such semantic representations to syntactic structures, applying morphological rules, and linearising the output as a string. Realisation in BT-45 is a relatively straightforward process, using an existing realisation engine [44]; hence, we shall not discuss it in what follows.

The microplanner recursively maps each of the nodes in a Document Plan ('messages' or events) to semantic representations (essentially predicate-argument structures with additional features), a simplified example of which is shown in Figure 8. This mapping involves three principal stages. *Lexicalisation* maps an event to a predicate (usually a verb). As described in Section 4.1.1, events have properties which specify the the entities that participate in the event (for example, the medical staff who perform an action, the drug given in a DRUG ADMINISTRATION event, etc). Lexicalisation rules also select these participants and map them to the thematic roles of the predicate (such as AGENT and RECIPIENT). The semantic content of these arguments is selected by the *Referring Expressions Generation* module based on the properties specified in the ontology. *Event Linking* performs a limited amount of aggregation on the resulting structures, linking them together based on some of the relations specified in the Document Plan. Of the links which are explicitly rendered, the most crucial is causality, whose expression depends on the (linguistic) type of the events. For example, if the linguistic representation for the caused event corresponds to a declarative clause, the Event Linking module adds a sentential adverbial such as *as a result*, giving rise to texts such as the following:

> By 14:27 there had been 2 successive desaturations down to 56. As a result, Fraction of Inspired Oxygen (FIO2) was set to 45%.

These three modules are mediated by a central *Discourse Manager*, which keeps a record of the events and entities mentioned, structuring the discourse into segments and determining their tense and aspect features based on the linguistic context. The expression of time is the most crucial task of this module. Given the importance-based heuristics used for document structure, the order of events in text does not necessarily reflect their temporal order. Since in the absence of further information, readers tend to interpret consecutive clauses as denoting temporally consecutive events

[86,131], the correct tenses (especially the perfect/non-perfect distinction) and adverbials are required to enable the reader to reconstruct the correct sequence. Tense is determined based on an implementation of Reichenbach's model of time [100]. In this model, whether an event is expressed using a perfect tense is determined by comparing its event time to a reference time. We adopted the strategy proposed in work by Webber [125] and Passonneau [92], whereby the reference time of an event is the event time of a salient, previously mentioned event, usually the most recently mentioned one in the discourse. If reference time is *after* the event time, the resulting clause has a perfect tense, as shown by the following fragment from Figure 6.

> At 17:15 FIO2 was lowered to 33%. TCPO2 had rapidly decreased to 8.8. Previously T1 had rapidly increased to 35.0.

In addition to tenses, all key events (which start each paragraph) are anchored in time by an explicit temporal reference (such as *at 17:15*). Events corresponding to long trends have temporal expressions that indicate their duration (e.g. *Over the next 24 minutes...* in paragraph 2 of Figure 6). Otherwise, relative temporal adverbials such as *previously* are used when there is potential ambiguity in the texts. For example, in the fragment cited immediately above, the increase in T1 occurred prior to the decrease in TCPO2, but this clause is already in the past perfect, giving rise to the possibility that the two events be interpreted as having occurred consecutively. The adverb *previously* is used to make it clear that the increase in T1 was the first of the two events to occur.

4.2. Evaluation of BT-45

BT-45 was evaluated in an off-ward experiment with clinicians, held between November 2007 and January 2008, in conditions that largely replicated those of the experiment of Law *et al.* [75]. Doctors and nurses were shown 45-minute *scenarios* – stretches of physiological and discrete data corresponding to a single patient – and asked to study them and decide, based on that information and some background text about the patient (written by experts), what clinical actions should be taken at the end of that period. However, while Law *et al.* manipulated a textual and a graphical presentation condition, our experiment included a third condition, in which data was presented as a textual summary automatically generated by BT-45. Thus, the experiment had the following principal aims: (a) to attempt to replicate the results of Law *et al.* on the utility of text versus graphics for clinical decision-making, focussing on the form of graphical presentations currently in use in the NICU; (b) to assess the feasibility of NLG technology as a decision-support tool, in relation to the kind of graphical presentation currently in use in the NICU; (c) to compare decision-making based on automatically generated text to that based on human-authored text. The experiment used historical NICU data to ensure that clinicians who participated in the experiment were not familiar with the patients, and could only make decisions based on the information presented. Full details of this experiment are reported by van der Meulen *et al.* [122].

4.2.1. Materials, design and procedure

Twenty four scenarios of approximately 45 minutes were prepared, in addition to two used to train participants on the experimental software. The data for each scenario consisted of physiological signals, numerical data and discrete event data. Scenarios were selected by a senior neonatal nurse and a consultant neonatologist. The main selection criterion was ensuring a balance among them in terms of the principal (or 'target') clinical action the patient's state called for. There were 8 such target actions, identified in advance of data preparation.

In addition to a target action, each scenario was also associated with a set of *apropriate* (that is, beneficial), *inappropriate* and *neutral* actions. The target action for a scenario was one of the appropriate actions[4]. There were 18 possible actions (including 'no action') that could be selected, different subsets of which were appropriate, inappropriate or neutral depending on the scenario.

The experiment was carried out using a modified version of the Time Series Workbench (TSW) [59]. An example of the interface is displayed in Figure 9. The main window contained the information in the experimental condition (text or graphics), while a separate panel presented the background text about the patient. Actions were selected in a different panel at the bottom of the screen.

Each scenario was presented to a participant in one of three conditions. Human-authored textual summaries (H) represented consensus summaries written by a consultant neonatologist and two experienced

[4]Note that, unlike the Law *et al.* dataset, target actions in our evaluation served only the primary purpose of guiding scenario selection. The evaluation metric used aggregated over appropriate actions.

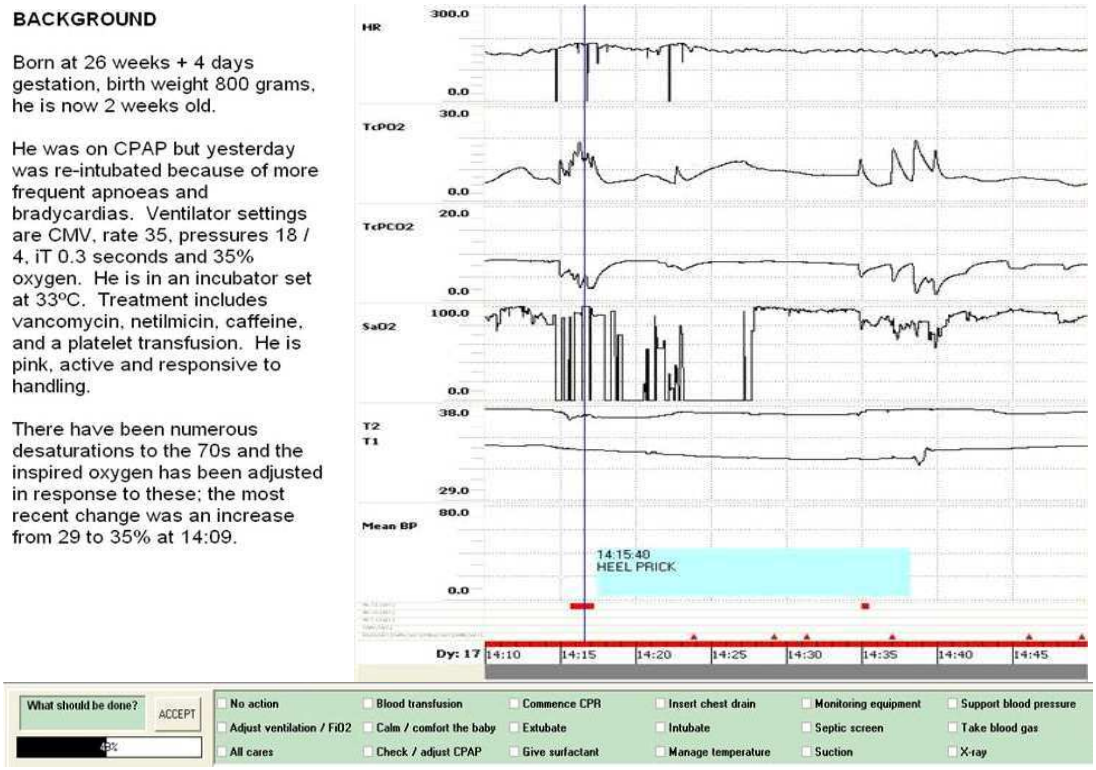

Fig. 9. Experimental software display in the graphical condition

neonatal nurses; an example is given above in Figure 5. The summaries were written to avoid excessive interpretation (such as statements to the effect that some physiological parameter was normal or too high). Graphical presentations (G) consisted of graphs of the physiological data with annotations of the discrete events, as shown in Figure 9. This display was intended to approximate the kind of visualisation provided by the current system in the NICU (see Figure 1). The discrete events were shown as coloured icons below the time-series plots, which a user could click on for more information. In order to avoid presentational overload, only the discrete events mentioned in the human-authored texts were presented as annotations to the visualisation. The computerized texts (C), an example of which is shown in Figure 6 above, were generated using BT-45 on a database containing all of the data (continuous and discrete) that was available to the human experts in writing their texts.

The test data used to generate the experimental summaries was not seen by the BT-45 developers in the course of building the actual system. Once the experimental texts were generated, they were only checked to ensure consistency in terminology with the H texts, and to identify any glaring errors. In fact, only three terms were manually altered in the experimental texts because they were of terminological inconsistency, and no errors needed correction.

Thirty-five staff members of the NICU at the Royal Infirmary of Edinburgh were recruited. They were divided by role (nurse or doctor) and by experience (junior: ≤ 4 years of experience; senior: ≥ 8 years of experience). There were 9 junior and senior doctors, 9 senior nurses, and 8 junior nurses.

Each participant saw all 24 scenarios (in addition to the two training scenarios) in one of the conditions. Presentation modality per scenario was counterbalanced using a Latin Square design, and order of presentation was randomised. Participants were not informed of how the texts in the H and C conditions had been produced. Each trial (presentation of single scenario) timed out after three minutes. This only occurred on three occasions.

4.2.2. Results

Table 1 displays the mean decision-making performance score, overall and within groups, as well as the time taken on average within each condition. Timings were estimated from the point where the data for a sce-

	J. Doctor	J. Nurse	S. Doctor	S. Nurse	Overall score	Mean time
G	.37 (.15)	.40 (.19)	.40 (.16)	.44 (.09)	.40 (.15)	73.16
H	.42 (.11)	.48 (.10)	.44 (.10)	.47 (.12)	.45 (.10)	77.23
C	.44 (.16)	.36 (.10)	.38 (.12)	.47 (.10)	.41 (.13)	78.81

Table 1

Mean decision-making performance score and standard deviations per group and overall, with reaction times in seconds

nario was initially presented, to the point where a participant made their first selection of an action. There were no significant differences between reaction times, either as a function of condition, or as a function of user group [122]. It is worth noting that on average, the time taken to process the data was well below the threshold of 3 minutes per scenario, suggesting that sufficient time was allotted.

A decision-making performance score was obtained by subtracting the proportion of inappropriate actions from the proportion of appropriate actions, that is:

$$\frac{|A_{as}|}{|A_a|} - \frac{|A_{is}|}{|A_i|} \quad (1)$$

where A is the set of 18 actions associated with a scenario, $A_a \subseteq A$ the set of appropriate actions, $A_{as} \subseteq A_a$ the set of appropriate actions actually selected; likewise for A_i the subset of inappropriate actions. Neutral actions were not counted in the score. The score ranges between -1 and 1. It is negative in case a greater proportion of inappropriate than appropriate actions are selected. This score was intended to account for the fact that the number of (in)appropriate actions was unevenly distributed over the 24 experimental scenarios. For alternative analyses of the decision-making performance in the experiment, we refer to van der Meulen *et al.* [122].

A 3 (Condition) x 4 (Group) by-subjects Analysis of Variance (ANOVA)[5] showed no main effect of Group, but an effect of Condition that approached significance ($F(2, 31) = 2.939, p = 0.06$). There was no interaction. Separate ANOVAs showed a difference between the G and H conditions ($F(1, 31) = 4.975, p < 0.05$) and the C and H conditions ($F(1, 31) = 5.266, p < 0.05$). Crucially, there was no significant difference between the G and C conditions. In a follow-up analysis, scenarios were grouped by the main target action. Here, a somewhat different pattern emerged: for five of the target actions, computer texts were were as effective as human texts, but were worse in the other three. A by-items ANOVA testing the effect of target action on difference in performance between H and C texts showed that this trend was significant ($F(1, 7) = 8.002, p < .001$). This result suggests that human texts may have contained information that was more relevant to some of the target actions than computer texts.

There are three aspects of these results that are worth emphasising. First, there are clear differences in the effectiveness of textual summaries compared to the graphical modality currently in use in the NICU. This emerges from the superiority of human-authored summaries relative to graphical presentations, effectively replicating the findings of Law *et al* [75]. Secondly, the difference is not so much in speed (no significant effects were found on reaction times), but in accuracy. Taken together, these findings suggest that a coherent summary can result in better decision making. It may also be the case that text offers no trade-offs in speed compared to visualisation of trend data, though this finding should be treated with caution, given the relatively short length of the scenarios used in the experiment (45 minutes). The third crucial point is that, as far as decision-making accuracy is concerned, BT-45 emerges as no worse than the visualisation techniques which are currently avalable in the NICU. This is encouraging because it suggests that NLG technology is viable as a decision-support tool, resulting in no detectable loss in decision-making accuracy relative to current techniques. On the other hand, the superiority of the H texts over the BT-45 summaries could be due to two factors. On the one hand, humans tend to be better at deciding what to include in a summary. Our follow-up analysis based on target actions suggests that this may be the case. On the other hand, they also tend to be better at expressing temporal and causal relations between events coherently. We turn to these issues in the next sub-section.

4.2.3. Discussion

The evaluation helps to focus attention on those aspects of the BT-45 texts that need to be improved[6].

[5] We used an ANOVA since this is the inferential method of choice for a mixed design which satisfies parametric assumptions.

[6] The discussion in this section is based in part on the outcome of a comparison between H and C texts carried out by two discourse

We argue that these challenges are primarily related to the narrative structure of the generated summaries. As discussed in Section 2.2, we use the term *narrative* to refer to a discourse dealing with a limited number of events which are temporally related, possibly aiming to emphasise a subset of these as more important or relevant.

Content Selection One of the problems evinced by the BT-45 texts is related to how content is selected by the Document Planning module. As noted in Section 4.1.4, the decision of whether to include an event or not is largely based on its *importance*, which is computed by the Data Interpretation module for each event. This heuristic ignores the possibility that the importance of some events may alter as a result of the context. A simple illustration of this is provided by the description of trends. In some cases, BT-45 mentioned two trends in the same channel, omitting an intervening event (deemed to have low importance) which would have played a role in making the summary coherent:

> TcPO2 suddenly decreased to 8.1. [...] TcPO2 suddenly decreased to 9.3.

In the above case, the first sharp decrease in TCPo2 is followed by a gradual increase, which then gives way to a second sharp decrease. The increase itself is relatively unimportant (e.g. because of its rate) if taken out of context, but becomes crucial if the second clause in this example is to be made sense of.

Another aspect of the content selection problem, which also has consequences for Data Interpretation and domain-specific reasoning, is raised by the observation that the performance of our experimental participants differed depending on the target action. This implies that a content selection strategy might benefit from a better-defined communicative goal, that is, a prior notion of what action the reader of a medical summary is expected to take based on the information imparted. In the BT-45 experiment, goal-driven summarisation was kept to a minimum. However, it is conceivable that guiding the generation process in this way would also improve the narrative quality of the output, since it would make it not so much a summary whose purpose is purely descriptive, as a story whose purpose is to emphasise certain events over others, particularly if these are related to low-probability outcomes (which may therefore be missed by the reader). As an example, BT-45 tended not to mention artefacts in the data, but mentioning artefacts can be useful if one of the goals is to get the reader to check the monitoring equipment. Indeed, the human summaries contain references to noise in the data on such scenarios. Clearly, the ability to select content in this way depends on the ability to identify possible outcomes in the first place, a task that places a larger burden on the reasoning component.

The temporal dimension As noted in Section 4.1.5, one of the challenges for microplanning in this domain arises from the fact that the order in which events are narrated does not reflect temporal order. The strategy which is used to relate events to each other in context occasionally gave rise to ambiguity, that is, to cases where the reader would find it difficult to reconstruct the precise order in which events occurred. An example is shown below:

> By 14:40 there had been 2 successive desaturations down to 68. Previously FIO2 had been raised to 32%. TCpo2 decreased to 5.0. T2 had suddenly increased to 33.9. Previously the SPO2 sensor had been re-sited.

In this passage, the increase in T2 (peripheral temperature) occurs prior to the previously mentioned event (the decrease in TCPO2). The re-siting of the SPO2 sensor occurs between these two events, and the microplanner attempts to disambiguate by using the temporal adverb *previously*, with limited success. A better temporal strategy would need to include a finer-grained model of the semantics of adverbials (e.g. [53]) and how these interact both with the type of event being described (for example whether an event is a state or whether it is telic; cf. [124,30,82] among others). In addition, a better model of the interpretation of events based on their tense and its interaction with other events in a discourse context is required. In particular, our strategy of setting the reference time of an event to the event time of the last-mentioned event in the discourse (cf. [92,125,131]) may have given rise to a lot of shifts in the narrative timeframe of a summary. A better strategy may be to fix the reference time within a passage (for example, to the time of the key event), and relate all events to this focal temporal anchor.

4.2.4. Methodological considerations

The evaluation experiment reported here fits well with a long-standing tradition in NLG of evaluating systems with their target users in task-based evaluations [104,65,114] (though such studies have been less prevalent among data-to-text systems). This dis-

analysts from the University of Edinburgh, Dr Andrew McKinlay and Dr Chris McVittie.

tinguishes such studies from related work in Human-Computer Interaction, where a methodology for evaluation with real users remains a challenge [94]. Nevertheless, such evaluations present certain methodological difficulties, principal among which is participant recruitment.

The number of participants in the BT-45 evaluation was relatively small. Given the relatively high variance among participants in the performance scores (see Table 1), this may have compromised the analysis. For example, the lack of main effect of participant group suggests that larger sample sizes are needed to identify such differences if they exist. Nevertheless, obtaining such large numbers of participants for an experiment of this kind is difficult. Unlike a laboratory-based experiment, the present study required medical professionals to take time off a busy schedule in order to participate. In a clinical environment, this raises several logistical difficulties, not least the necessity of replacing staff who are off-ward during the experiment.

In our current work on the new BabyTalk systems, particularly BT-NURSE (Section 5.1), we are planning to conduct evaluations on-ward, both to overcome these logistical difficulties, and to have the opportunity to evaluate systems not only with real users, but also within the real setting for which they are intended. Such an evaluation would also serve to assess the utility of NLG for decision support in a non-artificial setting using real-time data.

4.2.5. Interim summary

BT-45 provided an opportunity to test the feasibility of automatic summarisation in the NICU, comparing it with existing technologies for data presentation, and identifying some of the shortcomings of the technology. The results of the evaluation have prompted several follow-up analyses, focussing in particular on the differences between human and computer summaries that could account for the superiority of the former over the latter in supporting decision-making. We have discussed some of the main issues above, and are actively engaging in more detailed analyses of the experimental results [103,43]. In the rest of this paper, we describe how some of these insights are being taken into account in developing systems for different user groups in the neonatal context.

5. Generating summaries for different users

BT-45 generates textual summaries for nurses and doctors, to help them make decisions about appropriate treatments and other interventions. It was intended to be a decision support tool, and was evaluated as such. However, there are many other possible uses and users of textual summaries of NICU data, and indeed textual summaries of medical (and other) data more generally. For example, textual summaries can be used by clinical staff to support long-term planning (as opposed to short-term decision making), to check for errors in a patient record, and as an authoring aid for routine documents. Textual summaries can be used by management as a quality assurance tool, to help identify problems and successes. Last but not least, text summaries can be used by patients, to reduce stress and support informed decision making, to encourage behaviour change [104], and also to encourage friends and families to give appropriate support.

Previous research projects in medical Natural Language Generation have looked at many of these applications [17,12,56]. As observed in Section 3.3, these projects differ from the BabyTalk vision in that they have mostly focused on summarising small data sets which cover only a very limited portion of a medical record, and in particular do not include sensor data or detailed data about actions performed by medical staff.

We are now working on several other BabyTalk systems, for different user groups; these will give us insight on the utility in different contexts of textual summaries generated from large EMR data sets. Specifically we are working on BT-NURSE, which generates summaries for nurses to assist in shift handover and care planning; BT-FAMILY, which generates summaries for parents to keep them informed and reduce stress; and BT-CLAN, which generates summaries for friends and family, to encourage them to provide appropriate support. These systems are all currently under development. Another system, BT-DOC, which will generate summaries aimed for junior doctors, is planned for the near future.

In this section, we discuss the research questions which are being raised in the course of development of BT-NURSE, BT-CLAN and BT-FAMILY.

5.1. BT-NURSE*: Helping nurses plan care over a shift*

The goal of BT-NURSE is to generate textual summaries which will be included with nursing shift summaries (possibly after editing by a human nurse). Nurses at the Edinburgh NICU work 12 hour shifts, during which they continuously enter data. Shift summaries, of the sort exemplified in Figure 4(a), are produced by the electronic medical record system, which

Background
The baby was born at 24 weeks weighing 460g. He is 2 days old and in intensive care.

Respiration
Current Status
The baby is currently on CMV. Ventilator BiPAP rate (vent RR) is 55 breaths per minute. Pressures are 20/4. Inspired oxygen (FIO2) is 27%. Ventilator tidal volume is 1.5. The most recent blood gas was taken 11 minutes ago. Parameters are normal. Ph is 7.3. Concentration of carbon dioxide (CO2) is 5.72kPa. A suction was done. There were blood stained secretions and purulent secretions.

Events During the Shift
An ABG was taken at 23:09. There was evidence of respiratory acidosis. [...] The baby was moved from BiPAP to CMV. He had been intubated.
[...] Another ABG was taken in the early morning. There was evidence of respiratory acidosis. Ph was 7.18. CO2 increased to 8.74kPa.
Blood gas parameters had improved by 06:28. [...] The last blood gas was taken 11 minutes ago. Ph increased to 7.3. CO2 dropped to 5.72kPa.

Potential Problems
Purulent secretions during shift suggest risk of infection.

Fig. 10. Excerpt from a summary automatically generated by the current version of BT-NURSE

collates data entered throughout the shift under headings related to the physiological system to which they are relevant. In addition, nurses add further information by typing it directly into the system at the end of their shift. Apart from the shift summary document, the incoming and outgoing nurse also orally discuss the babies they are caring for.

Shift summaries are used by the incoming nurse to help plan how to best care for the baby. Ideally they should give a summary of the baby's history and current status and describe how the baby has reacted to previous interventions. Such an 'ideal' summary was exemplified earlier in Figure 4(b). Figure 10 shows an excerpt of a shift summary generated by the current version of BT-NURSE, which reports exclusively on events related to a baby's respiratory condition. The summary illustrates how we are trying to approach the ideal, overcoming some of the shortcomings of current presentation formats identified in Section 2.1: unlike the BT-45 texts, it goes beyond the description of a series of events, giving backgound and current state information, as well as pointing out potential problems. Among the improvements envisaged for BT-NURSE are better text planning to highlight temporal and causal relationships between events, and coverage of other physiological systems besides respiration. The ultimate aim of BT-NURSE summaries is to help both outgoing and incoming nurses. In the case of the former, the idea is to allow them to decide on whether to include any of the generated text into their own written notes, possibly with changes or revisions, or use BT-NURE summaries in their entirety. In the case of incoming nurses, integrative summaries of this kind can help to achieve better care planning.

From a scientific perspective, BT-NURSE raises a number of important challenges, including the following.

Identifying the time of events From a data analysis perspective, perhaps the major new challenge in BT-NURSE is identifying when events occurred. The input data to BT-45 included fairly accurate information about the timing of actions and events; this is because it was taken from a special data set created for research purposes. The input data to BT-NURSE includes much less accurate timings; in many cases we know an event occurred in a 60-minute interval but not where in that interval. This problem occurs with much of the manually entered data, because observations, events and actions, such as those in Figure 3, are typically logged after they have taken place. We are working on using the signal data to refine timings, by looking for patterns in signals which can be correlated to the relevant actions and events; some further aspects of this challenge are discussed in Section 6.1.4. From the linguistic point of view, the BT-NURSE microplanner has been designed to use vague temporal expressions (such as *in the early morning* in Figure 10) rather than precise times such as *at 06:00*. In this way, events are situated in time in an approximate manner, reflecting the temporal uncertainty.

Reasoning about plans and goals From a data interpretation and reasoning perspective, a major new challenge in BT-NURSE is making inferences about nursing goals and plans. BT-45 described data but did not make higher-level inferences about diagnoses and intervention, which probably hurt its evaluation performance because, as discussed in Section 4.2, it did not estimate the importance of events in context, leading it to ignore issues such as when artifacts should be mentioned. BT-NURSE already makes inferences about potential problems and care actions, which are communicated in its texts; our medical collaborators believe this is essential in assisting junior nurses to do better care planning. Hence BT-NURSE will have to do even more higher-

level reasoning, which in principle could involve nursing protocols (although we may not get this far).

Reasoning about goals also has an impact on the narrative structure of the generated summaries. Some aspects of this emerged from our comparison of the two summaries in Section 2.1. There, we argued that the summary in Figure 4(b) was more goal-driven than the one in Figure 4(a), emphasising certain events and relationships which would draw the attention of a nurse to the principal issues in drawing up a care plan.

Consistency and cohesion between overlapping texts As shown in Figure 10, BT-NURSE texts are divided roughly as in the corpus summary exemplified in Figure 4(b). Many of the sub-parts of the text have overlapping content, and from a document planning perspective, a major challenge in BT-NURSE is managing this overlap: for example, deciding when information needs to be repeated and when a simple reference to a previous portion of the text suffices, and also ensuring that the different parts are consistent when discussing similar facts. In short, some of the continuity problems identified above for BT-45 become more crucial in the context of BT-NURSE.

As an example of how continuity and cohesion needs to be better managed, consider the mention of *the most recent blood gas* in the *Current Status* section of Figure 10. This test is described here because it has immediate bearing on an assessment of the baby's current state; however, it is also mentioned again under *Events During the Shift*. Since this is the last such test undertaken in the relevant period prior to summary generation, it is now referred to as *the last blood gas*. However, the text does not make it clear that this is the same test that was mentioned at the beginning of the summary, giving rise to potential confusion. This example illustrates the non-trivial decisions that need to be taken, involving when to introduce certain facts and when to mention them again in a later portion of the text, as a function of the *purpose* of the discourse (for example, whether a particular section is intended to give an overview of a current state of affairs, or whether it is narrating a sequence of events).

Reference Given the overlapping structure of the various sub-sections in a shift summary, BT-NURSE texts will need to include references to entities, actions, analysis, etc made elsewhere in the text, making reference a much more central issue than it was in BT-45. Our earlier example of the two references to the same blood gas sample is also a good illustration of the problem of reference and anaphora: the initial reference says *most recent*, which makes sense given that the purpose of that section is to describe current status, but how should the second reference be rendered? The question is complicated by the fact that the text mentions several blood gas samples in the *Events During the Shift* section (e.g. the sentences starting with *An/other ABG was taken...*) before the most recent one is mentioned for the second time. One possibility is to use a document-deictic reference [90], such as *the blood gas mentioned under* Current Status. It seems that using *the last*, as the system currently does for the second reference, does not make the link sufficiently clear.

The question of anaphoric reference has been treated to some extent in the generation literature, with the primary focus on pronouns and reduced descriptions [48,67,71]. However, a strategy for repeated references in BT-NURSE must interact with the outcomes of text planning [68]. As our example illustrates, there is an interaction between content selection for referring expressions and the goals of particular segments of text, something which has not received much attention in the NLG literature (exceptions to this trend include the work of McCoy [78] and Jordan [62]).

5.1.1. Prospects for evaluation

Our planned evaluation for BT-NURSE is intended to overcome some of the methodological difficulties outlined in Section 4.2.4. In particular, we are planning to deploy BT-NURSE live within the NICU, over a period of six months. Hence, unlike the BT-45 evaluation, which was based on archived patient data, the evaluation of BT-NURSE will be based on live data about current babies, with nurses looking at BT-NURSE texts about babies currently under their care. Although deploying BT-NURSE on the ward does not raise any novel research issues beyond those related to developing the core generation technology, it does require addressing a number of software engineering issues which were much less important in BT-45. More importantly, it raises questions related to the design of the study.

Carrying it out in a real setting implies less control over the sample of participants. The hectic rhythm of activity in the NICU also imposes limitations on what participants should be required to do. Our plan is to perform different evaluations with outgoing and incoming nurses. The former will be shown a summary generated about the baby they are caring for, and asked to rate each segment of the summary in terms of its utility and correctness. The same ratings will be elicited for the summary as a whole. They will also be free to

comment on any aspect of the summary they feel is relevant. The purpose of this part of the evaluation is thus to assess to what extent BT-NURSE is capable of facilitating the work of an outgoing nurse in writing a shift summary. The task of an incoming nurse, on the other hand, is to plan the shift based on the information in a summary. Hence, incoming nurses will be asked to first formulate a care plan based on the usual information sources (including the oral handover by the outgoing nurse), then read a BT-NURSE summary and highlight any changes in their plan as a result. Since care plans are not typically written, we envisage a 'think-aloud' protocol for this part of the study.

5.2. BT-FAMILY*: Informing stressed parents about the status of their baby*

BT-FAMILY builds on an earlier parent information system developed at Edinburgh, called BabyLink [39]. Its purpose is to generate informative summaries for parents of pre-term babies, to keep them up to date about how their child is doing. Having a child in neonatal care can potentially cause a considerable amount of stress and anxiety for the parents. At such a time effective communication between medical staff and parents is needed not only to inform, but also to reassure. Whilst medical staff are usually very willing to talk to parents, the provision of additional information through person-to-person communication may not be fully adequate, given the time constraints of both staff and parents (who often have other family matters to attend to, including other children). Therefore BT-FAMILY is being designed to provide parents with regular information summaries about the condition of their child. Additionally, BT-FAMILY will explore how such information must be tailored so that it is takes the emotionally sensitive state of the parent into account and avoids the possibility of creating any additional distress. Information summaries created by the system will be presented using a Web-based interface, which can be accessed remotely, using a password.

From a scientific perspective, an obvious challenge in BT-FAMILY is to present information in a suitable way to non-experts. This topic has been addressed in a number of previous research projects, where reader expertise has been used to inform decisions about lexical choice and terminology [80] and document structure [91]. In a similar vein, Williams [127] varied a number of linguistic parameters based on literacy level. We are doing some work along these lines (one complication is that parents often say they prefer 'medicalese' texts, even if they don't understand them [63]).

However, the primary research focus in BT-FAMILY is on how information with high emotional impact (i.e., updates on the the status of a very sick baby) should be communicated textually to a reader (the parent) who is probably under a considerable amount of stress. Previous work on Affective NLG has tentatively shown the possibility of dynamically generating variations of a text based on constraints related to its emotional impact [55,27,26]. However, this work has tended to stop short of empirical evaluation of its claims [9]. More recent work has shown that textual variations can have an emotional impact on the reader [123], but this has primarily focussed on artificial experimental contexts, where it is unclear how much subjects care about the information they are given. BT-FAMILY, in contrast, communicates real information of high emotional impact to people who deeply care about it.

BT-FAMILY's first task is to estimate the stress level of the parent reading the text. The predictive model we are creating is based on previous research, which has shown (not surprisingly) that the best predictor of stress level is how well the baby is doing, and whether the baby is getting better or worse [28]. The second task is to adapt generated text according to predicted parental stress. Knowledge acquisition studies (discussions with experts, analysis of corpus texts, limited interaction with parents) has suggested a number of possibilities. One suggestion is simply to say less to people who are stressed, because stress reduces the ability to absorb information; a related principle is to avoid numbers, acronyms, and other technical details if the reader is stressed. Another idea is to add reassurances and explanations for moderately stressed readers (these might not be appropriate for highly-stressed readers as they make texts longer). We will implement and explore these ideas over the next year.

From a practical perspective, one of the difficulties in carrying out this research is its ethical dimension. For example, detailed knowledge acquisition activities with parents of NICU babies is necessarily restricted because of concerns about the adverse effect such interaction might have on people who were already under severe stress. For similar reasons it will probably not be possible to evaluate BT-FAMILY with parents of current NICU babies, although we will be able to do so with parents whose babies were in NICU but are no longer there. We suspect similar ethical considerations are likely to constrain other research into communicating emotionally sensitive information to stressed individuals.

5.3. BT-CLAN: *Soliciting support from friends and family*

The goal of BT-CLAN is twofold: (a) to reduce the communication burden of parents and (b) to encourage friends and family (the social network) to provide appropriate emotional and practical support to the parents. Parents may find it overly time-consuming to provide updates about the baby to their social network, when they are already busy with their sick infant. Yet, members of the social network need to know how the baby is and what assistance parents need if they are to give appropriate support.

BT-CLAN is informed by findings in evolutionary anthropology which suggest that there is a consistent hierarchical structure to human social relationships [32]. Individuals typically have 3 to 5 people who are very close to them, 12 to 20 people who are close but less intimate, 30 to 50 people with whom they associate on a regular basis, and a larger group (typically in the hundreds) who are merely acquaintances. Our user studies have indicated that people in each group want different kinds of information at different levels of detail. One of the notable aspects of these findings is that network members tend to want more information about the parents than about the baby [83]. Members of the closest group are likely to want a regular high-level summary of how the baby and the parents are doing. In contrast, members of the less close groups want less information, less often. BT-CLAN asks parents to place friends and relatives into appropriate groups, and to give information about themselves and their support needs. The user model is then applied to define what information members in each group should receive.

The system, which is currently under development, will be under the control of the parents, who can specify what information different friends and family should receive, and also what help they would like to get. Based on this input, BT-CLAN sends automatic updates to network members about the baby and the parents, and also about ways in which network members can help, from practical tasks such as providing childcare and help with laundry, to emotional support. The aim of BT-CLAN is thus to relieve parents of the burden of responding to individual information requests.

One of the challenges in BT-CLAN is to allow for 'prevarication' which may occur when parents consciously seek to give false information about the state of their baby to certain network members. This may be motivated by concerns about the wellbeing of the members themselves. For example the parents may not want an elderly relative with heart problems to be informed that the baby is getting worse. Similarly, misinformation may be a defensive measure on the part of the parents, for example when they find a network member's interest in the baby overly intrusive, and may not want to give away too much information. In such situations, parents may want such a network member to be told that the baby is making progress, even if this is not in fact the case. To the best of our knowledge, very little research as been done on the automated generation of 'misinformation' on behalf of an individual. One question that arises is when such misinformation is appropriate, and when it is better to simply hide information; another issue is how to maintain consistency across several reports sent to the recipient, when the latter is being systematically misinformed.

6. Discussion

In this section, we discuss some of the broader challenges raised by the work described in the previous two sections. We group these under the two headings outlined in Section 1.

6.1. *Data-to-Text Issues and challenges*

6.1.1. *Narrative Structure: Building a Coherent Story*

Throughout the foregoing discussion, we have highlighted several features of narrative discourse which the family of systems under construction will need to address. It is worth summarising these features and setting out the challenges they pose to NLG systems, against the background of the BT-45 evaluation and our current work.

Goal-driven content selection Labov's work on the narratives that people produce as part of everyday verbal interaction has emphasised the fact that these discourses generally focus on a small set of events which are linked and which the narrator is including because they are germane to some overall communicative goal [73,74]. We have argued that the same charactistics are found in nurse shift summaries (see Section 2.2), and that the lack of goal-driven content selection was one of the weaknesses of BT-45 (Section 4.2). This kind of goal-driven communication goes beyond that typically investigated in NLG (e.g. [55]) or in computational creativity models of story generation [45]. In addition to the *a priori* goals typically handled by these models, our scenarios require subsidiary goals to be identified by reasoning with data, in order to find the

possible courses of action that the reader may need to take. A generation procedure capable of identifying such events, and of producing a narrative that will focus the reader's attention on some courses of action, would be very much in line with a view of narrative as an 'instruction' to the reader to actively construct a mental model of a situation which can then be used to support further reasoning [47,133].

Temporal grounding Though psycholinguistic work on narrative has emphasised the centrality of the temporal dimension [133,131], relatively little work has been done on the expression of time in NLG [86], as compared to Natural Language Understanding [125, 82,76]. The crucial challenge here is to be able to generate a reasonably long text, possibly with several narrative time shifts and following relatively fixed document structure conventions, from which the reader can nevertheless relate different events to eachother in time. From a linguistic point of view, this requires a production-oriented model of time in language which accounts not only for the correct use of tense and aspect, but also its interaction with adverbial modification and event anaphora (that is, nominal reference to events mentioned previously in the discourse). Moreover, tenses interact with discourse mode [16,111]. In particular, whether a section of a summary is about a patient's current state, or whether it is talking about events during a shift, will influence the choice of tenses and the way events are related to eachother or to the time of utterance.

Causality Readers have been shown to make inferences about causality and motivation continuously while reading a narrative text [47]. In addition to the burden it places on the reasoning component, causality also raises linguistic questions, especially relating to how it should be expressed, and whether it needs to be made explicit given the reader's level of expertise. The temporal contiguity of two events may suffice for an expert to infer a causal relationship, while less expert readers may need an explicit indicator. From a linguistic point of view, the expression of causality will of course interact with the expression of time.

Referential coherence In addition to temporal coherence, there is also plenty of psycholinguistic evidence that keeping track of entities in discourse via reference is another central dimension in narrative understanding [133]. We have already discussed some of the requirements that our current systems impose on the generation of referring expressions (Section 5.1).

Satisfying all these requirements imples many developments at every level of the data-to-text architecture. We discuss the consequences in the following sub-sections.

6.1.2. Context-sensitive Document Planning

Generating coherent narratives requires a more sophisticated model of Document Planning than that used in BT-45, whose heuristics were largely based on event importance, with little consideration of how importance changes as a function of context. For example, the scenario summarised in Figure 6 had the management of temperature as one of its main target actions. In this case, the Human and Graphical conditions gave better results than the BT-45 text, which contains only three references to T1 (compared to 5 in the human text, which also include a reference to peripheral temperature, T2). The first two are at the beginning *Core Temperature (T1) = 36.9. Peripheral Temperature (T2) = 36.6.* and the last at the very end *Previously T1 had rapidly increased to 35.0*. There are three main issues that a context-aware document planner must address, and which go beyond a purely bottom-up strategy.

Continuity As discussed in Section 4.2, BT-45 sometimes omitted 'unimportant' events which then gave rise to apparent inconsistencies. Thus, the statement that *T1 had rapidly increased to 35.0* does not make sense given the previously mentioned value of 36.9; an intermediate fall in temperature needed to be reported as well. Interestingly, while this shortcoming was criticised by participants in our evaluation experiment, some of the apparent continuity problems in the human texts were not. This suggests that some kinds of discontinuity can be tolerated as long as the global picture of the parameter evolution over time is maintained.

Selection of events from related sources The BT-45 document planner often did not select events from sources which, from an expert's point of view, need to be described in tandem. For example, the text under discussion did often mentioned T1 on its own, whereas the human texts always report T1 and T2 together. The same strategy is also evident in the nurse shift summary of Figure 4(b), where experts tend to group events related to the same physiological systems (e.g. respiration, thermo-regulation, nutrition etc.) and events derived from the same sensor. Thus, a Document Planner needs to make use of domain knowledge to link different parameters and structure different subsections.

Going beyond importance BT-45's importance-based content selection was coupled with thresholds to control text length, with the importance of each event computed in a independent manner. Not only does this ignore context, but it also fails to consider whether an event in itself represents a problem, or whether mentioning it will contribute to the main goal of getting a reader to attend to a particular set of tasks. This content selection strategy may therefore be better suited to summarisation rather than decision support. For example, in Figures 5 and 6, both summaries use a large amount of text to describe the re-intubation period (period from 17:00 to 17:15) but the human one always reported the temperatures in each paragraph while BT-45 was focused on events of very high importance and did not have enough space to add less important events (re-intubation is much more important than variation in temperature). More relevant content can only be achieved by a mechanism that addresses the global context. For example, if temperatures are below the normal value during much of the period being summarised, temperature-related events should be given more importance, whereas a successful re-intubation should be summarised in less detail.

All these issues point towards a strategy whereby the Document Planner performs a high-level overview or assessment, with further details being included to support the initial appraisal. Whether this strategy is applicable in a generic way is unclear; it relates to the general issue of how a data-to-text system could be linked to a medical diagnosis system.

6.1.3. Knowledge acquisition and modelling

It should be clear from the architecture of the different systems that the ontology is an essential component in our systems, linking data to its linguistic expression and supporting reasoning. Building a domain specific ontology is a time consuming, challenging and frustrating task (it is very rare that experts agree on a unique ontology). In BT-45, we chose to build an ontology from the ground up, but this may not be a satisfactory strategy for medium to large scale systems which aim for more general applicability. For example, BT-45 manipulated around 550 concepts. Our first attempt to fit the ontology to the actual clinical database using synchronization with UMLS led to about 1900 concepts and 70 properties. However, the end product needed to be refined due to coherence problems between UMLS and the application domain (e.g. UMLS does not have a DESATURATION concept and is not always consistent [21]).

Another fundamental difference is that the Neonate database used for the development of BT-45 contains intervals during which events *are happening* whereas the actual clinical databases contain information about events that have *happened*. Thus, moving from BT-45 to the systems described in Section 5 has involved a transition from an emphasis on actions (e.g. being intubated) to an emphasis on states (e.g. the baby is intubated, and this has occurred at some time within the preceding period). The expert rules acquired for BT-45 need to be adapted to deal with this new representation since events in the NICU database cannot be easily represented as time intervals with an accurate start and end time. Moreover, in line with our increased emphasis on narrative, future systems will need to perform high-level reasoning about problems, symptoms, actions and goals, using these to detect events of interest and reason about relations between them. For instance, if a ventilation parameter has been raised following a decrease in saturation, but this did not resolve the problem of falling oxygen levels, then the text should emphasise the fact that a goal did not succeed. Of course, such reasoning requires an exponential increase in the number of rules; we are investigating how the OWL ontology we are using could be exploited to carry more of the burden of reasoning than was the case in BT-45. An even greater reasoning challenge, which harks back to the classic problem of open versus closed-world reasoning, is to infer absence of events or negation. These feature quite frequently in free-text nursing notes in our database, which include observations such as *she is on no treatment for hypotension* or *blood was taken for an arterial blood gas 12 minutes ago but the result is not available*. In addition to the knowledge needed to generate such messages, there is also the problem of ascertaining that events have not been recorded because they have indeed not occurred (rather than through oversight, or because they were noted in free text elsewhere in the database). Thus, it may be safer for the system to state that no record of a particular event has been found, rather than stating categorically that such an event did not occur. The tendency to give false alarms or false diagnoses is one of the main reasons behind the success or failure of decision support systems, since it determines whether they are perceived as unreliable.

Finally, all these factors need to find expression in natural language. Microplanning in BT-45 assumed a relatively straightforward mapping between ontology concepts and linguistic expressions. However, more sophisticated linguistic interpretation is needed, using

a more fine-grained temporal ontology for natural language which may diverge from the core domain ontology. As an example, consider the identification of event times, which depends in part on whether an event is of relatively long duration, and on whether it culminates in some result. Thus, in one scenario, a BT-45 text stated that *After 6 attempts, at 14:17 a peripheral venous line was inserted successfully*. But 14:17 was actually the time of the first attempt and not the time at which the venous line was finally inserted. The microplanner therefore needs to reason about the internal structure of events and their telicity. These distinctions underlie much of the discussion of the lexical aspect of events in the formal semantic literature [124,30,82].

6.1.4. Temporal reasoning: Expressing vague or uncertain events

Most of the data handled by the data-to-text systems has a temporal nature (as opposed to data consisting of static characteristics such as gender, parental relationship etc.), but is stored in classical databases due to the absence of dedicated temporal databases on the market [118]. This situation, as well as the fact that most of the data is recorded manually, leads to several problems of temporal representation and uncertainty, with consequences for reasoning and linguistic expression.

In the data that our systems process, timestamps usually refer not to the time of occurrence of an event, but to the transaction time (i.e. the time at which an observation was logged). The difference between the two can be very large and the problem is exacerbated when events are reported at different temporal granularities. Free text data often uses coarse granularity (e.g. *parent visited yesterday*), while physiological signals are sampled at the frequency of once per second (e.g. bradycardia at 17:32:25). Imprecision implies that the temporal order of events cannot always be established, and makes reasoning more complex. While there has been some work to extend classical temporal reasoning theories, such as the Event Calculus, to deal with data at different time granularities [20], the formalisms that handle temporal uncertainty using domain knowledge either rely on fuzzy sets [31,7] or on probabilities [50,108]. To the best of our knowledge, these formalisms are rarely deployed in real-world scenarios and do not provide solutions to the integration of domain knowledge in order to constrain temporal information. One approach currently being investigated involves the use of different sources of information (databases, signals, free-texts) to detect the most possible occurrence time of inaccurately recorded events against a predefined temporal model [41].

Another major difficulty concerns abstraction or interpretation. Although they are rarely separated in the literature [113,130], we distinguish *temporal abstraction*, which represents a set of events into a single more abstract description, from *temporal interpretation*, which infers associations between events and relies much more on domain knowledge. Broadly speaking, temporal abstraction produces information that is nominal in nature (e.g., a sequence of events) whereas temporal interpretation yields representations that are more akin to propositions (e.g., A is linked to B, A is successful). The problem is that imprecision at the data level percolates up to the knowledge level, where inferences need to be made with varying degrees of certainty. For example, given a change in the ventilator (V) recorded after the occurrence of a desaturation (D) detected on the signals, and the knowledge that an increase in ventilation is usually performed to counteract a desaturation, a reasoning system could derive, as its most possible hypothesis, that V was caused by D, though there is no certainty of this given the uncertainty in the occurrence time of D and the fact that a desaturation can be resolved spontaneously. We are currently investigating techniques relying on possibility theory [31] which have shown interesting results in ICU diagnosis with accurately timestamped data [89].

It should be clear from the foregoing discussion that though domain knowledge reduces uncertainty, it does not eliminate it. Hence, uncertainty and vagueness need to be communicated. Though recent work in Natural Language Understanding has addressed the task of extracting temporal relations in medical free-texts [107], less attention has been paid to the expression of uncertainty in NLG systems using mechanisms such as vague predicates [120] and modals [69]. The use of some of these mechanisms is exemplified in the following free text fragment:

> Frequent and sometimes profound desaturations with bradycardia. Cultured and started on gentamicin to cover possible sepsis.

This text uses vague temporal expressions and relations (*frequent*, *sometimes*, *with*), expressions of possibility (*possible*) and a vague term to describe magnitude (*profound*).

Of course, as technology improves and data collection becomes increasingly automated, uncertainty in data is expected to reduce. However, language in the medical domain is replete with vague descriptions (e.g. *the patient is doing well*, *the baby is blue*), and uncertainty will always arise from reasoning with such data.

Hence, a unified theory that handles both vagueness and uncertainty is required. Our work on the BabyTalk systems has led to some exploration of these issues [95].

6.1.5. *Extraction of Information from Raw Data and Free Texts*

The data available for data-to-text technology is often raw or unstructured, whether it is in the form of time series, images, video or free text. Thus, techniques from signal processing and Natural Language Understanding are relevant for this kind of technology.

Signal processing in BT-45 was relatively simple (see Section 4.1.2). More sophisticated approaches relying on probabilistic modelling, such as switching Kalman filters [99,4] would give more accurate results. However, as pointed out by Aleks [4], these methods are quite time-consuming and the number of tests performed increases dramatically with the number of searched patterns. Since summarisation systems require a lower degree of accuracy than diagnosis systems, a trade-off in favour of efficiency may be desirable. Nevertheless, future BabyTalk systems require improved methods to detect artifacts, especially because some artifacts have a signature which comes close to that of actual patterns of interest, increasing the likelihood of false classifications. Sometimes noise occurs *during* a period where a real event is in progress; hence the problem is not only to filter artifactual values but also to reconstruct the actual values based on surrounding context. This problem is far from being solved and comparative studies have not succeeded in identifying a single method of choice [93].

In BT-45, no free-text was used to produce the final texts. However, these notes contain very valuable information, including the justification for medication, descriptions of medical or surgical interventions, information about a baby's parents, etc. These free texts are often highly unstructured, with lots of abbreviations, grammatical errors and local jargon. We are currently employing Information Extraction (IE) techniques [107] to automatically extract information from these free-text notes. One of the challenges is to be able to identify redundant information, which is already available elsewhere in the database. This is non-trivial because free text, unlike database entries, is not timestamped.

Finally, our systems do not employ any visual information such as medical imaging or video recordings on the ward. However, doctors and nurses have repeatedly told us that they get a lot of useful information by visually observing babies. We intend to discuss with colleagues in the computer vision area the feasibility of bringing such information on the ward.

6.2. *Decision and information support issues and challenges*

6.2.1. *Textual summaries as a gateway to large multimodal databases*

All the BabyTalk systems are designed to generate textual descriptions of continuous and discrete (numeric and symbolic) multimodal data. Although our evaluation of BT-45 explicitly compared text *against* graphical presentation, we are *not* claiming that data-to-text technology has a definite superiority over visual approaches such as those reviewed in Section 3.1. Many visualisation techniques address issues that are complementary to those addressed in this paper. For example, the knowledge-based KNAVE II system [77] enables the visual exploration of laboratory data over a period of time and their interpretations (e.g. normal or abnormal values, bone marrow toxicity, etc). However, while these systems are very effective with experienced users, junior clinicians may not benefit as much from such presentations. This observation was confirmed by our evaluation of BT-45, which showed that junior clinicians performed better with textual presentations whereas senior clinicians' decisions are less subject to the presentation modality. Moreover, different types of data may be better suited to visualization or textual summarisation. Thus offering mutlimodal presentation (texts, graphs, images, sounds, videos of baby, etc.) would allow the user to choose the one which best fits her level of expertise and her own preferences. Ideally the presentation modalities could be linked with cross-references and otherwise integrated [5]. But whatever the chosen presentation, the data-to-text technology could be used to provide an initial overview (that is, a *gateway*) of the data to guide the user to the most salient information in the database. As pointed out by André [5], multimedia presentations follow similar structuring principles to those found in text, thus the extension of our systems to deal with multimodal presentation would not require the modification of the core architecture of the systems.

Another important improvement would be to make the BT-Systems interactive by including hyperlinks [88] which users could click on to generate a more detailed summary about particular events or to pop up the graphical representation of the raw data related to specific periods. Once again, KNAVE II [77] offers a good example, in that it enables users to query patient

databases to retrieve raw and abstracted low frequency data and display this on the screen for more accurate decision making in the oncology domain.

6.2.2. Description vs Recommendation Systems

Clinical decision support systems can be roughly divided into 3 families according to their output type: (a) systems that emphasize particular data (BT-45 is one of these) or enable interactive exploration (such as KNAVE II); (b) diagnosis systems (MYCIN being a classic example) ; and (c) recommender systems such as those which handle computerised clinical guidelines. Automatic recommendation or diagnosis is still a sensitive subject in medical practice and, with a few exceptions [29], has resisted commercialisation and widespread adoption on the ward. For example, recent work on the SONOCONSULT system [98] in the sonography domain, showed that clinicians preferred its data exploration service and tended not to favour its diagnosing service even though they acknowledged its correctness. However, the boundary between diagnosis, recommendation and summarisation may be not completely crisp. Systems like BT-45 could be extended to increase the amount of advice given in some parts of the generated text. This is one of the issues being considered in the development of BT-NURSE. As an example, Figure 11 illustrates how the same piece of data could be include recommendation and diagnosis to different degrees. With 100% diagnosis we would have a classical diagnosis system (which would state, for example, that the patient has pneumonia); 100% recommendation would yield a system akin to computerised guidelines (which would, for example, recommend a nurse to check urine output). Using NLG to fine tune the amount of advice in the output seems a very promising approach to overcome the current limitations of diagnosis and recommender systems.

6.2.3. Data-to-text technology to enhance transnational e-health

There is a growing interest in making health systems and patient data accessible from any country. For instance, the EU is considering health data sharing in its public policy agenda and conducts evaluations of its members' health systems, taking e-health infrastructures into account [52]. Large databases are being designed to gather terabytes of data (by hospitals, GPs, etc.) that will enable more accurate diagnosis of an individual patient's condition, the ability to mine population data to track health trends, etc. Access to these large databases is required because in future, patients will cross borders more often in search of cheaper or more specialised care. While there is currently a drive to design such databases, with adequate data formats and exchange protocols, solutions to make this data readable from any country in any language will eventually also needed. In the long run, data-to-text technology may provide part of the answer, particularly since it can be extended in principle to a multilingual setup.

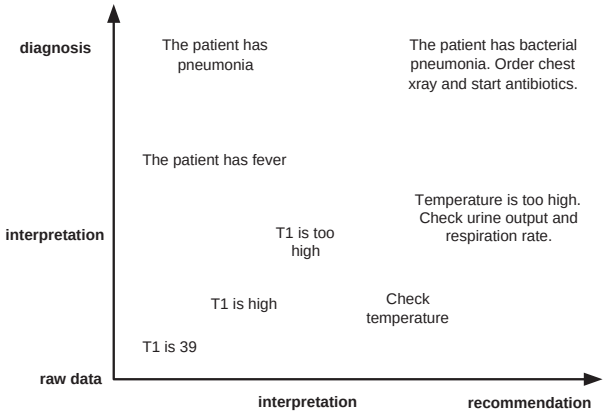

Fig. 11. Example of decision support output from the same data at variable degrees of recommendation and diagnosis.

Multilinguality will mainly impact microplanning and realisation, and we are investigating the extent to which these tasks can be defined in a generic fashion while externalising language-specific resources. The NLG community has in recent years seen moves towards making non-domain specific resources available, particularly where realisation components are concerned [33,126]. However, these are seldom explicitly designed to support multilingual applications, with some exceptions [8]. The prospects for generic microplanning components seem more daunting, not least because while there is consensus on what the primary tasks of this component are [102], their architectural organisation tends to differ from system to system [81]. Nor is it clear to what extent the nature of tasks such as aggregation is dependent on the target language. To our knowledge, there has been no systematic investigation of these issues from the point of view of multilinguality.

7. Conclusions

This paper has described recent and ongoing work on data-to-text systems in the domain of Neonatal Intensive Care. We have described the implementation and evaluation of one system, BT-45, using this as a starting point to identify the major challenges which

are currently being faced in generating medical summaries for different kinds of users.

Our approach to these challenges has relevance beyond the medical domain. As data collection techniques improve, so do demands for more accurate and informed decision making, but this often gives rise to serious problems of information overload. We have argued that the automatic generation of textual summaries constitutes a viable solution to this problem, especially if it is exploited in conjunction with other ways of handling data, such as visualisation. Moreover, our evaluations so far have suggested that, where effectiveness for decision support is concerned, the technology is on a par with existing methods. However, we have also pointed out several directions for future research. A central theme that emerges from our work is the importance of narrative as the overarching strategy for the summarisation of data in which time plays a central role, with important consequences at every level of the architecture of a system.

Acknowledgements

The authors would like to thank the other members of the BabyTalk team (Yvonne Freer, Felix Gao, Robert Logie, Neil McIntosh, Marian van der Meulen, Cindy Sykes and David Westwater) for all their help, and also the doctors and nurses who participated in the evaluation. Thanks are also due to Andy McKinlay and Chris McVittie for their help in the BT-45 evaluation. This work was supported by UK Engineering and Physical Sciences Research Council (EPSRC), under grants EP/D049520/1 and EP/D05057X/1.

8. Appendix: Glossary of medical terms used in the article

ABG An Arterial Blood Gas, that is, a blood gas where the sample is taken from an artery.

Apnoea Temporary cessation of breathing.

Arterial line/catheter Narrow tube inserted into an artery for measuring blood pressure or for obtaining a blood sample.

Blood gas A blood test carried out to determine pH levels in the blood, as well as oxygen, carbon dioxide and bicarbonate levels.

Bradycardia A brief episode of low heart rate.

Core temperature (T1) Temperature at the core of the body, typically measured at the chest region.

Desaturation Fall in oxygen saturation.

Extubation Action of removing an endotracheal tube from the baby's trachea.

FIO2 Fraction of inspired of oxygen setting on the ventilator.

HR Heart rate from electrocardiogram leads or arterial catheter.

ICU Intensive Care Unit.

IV line See peripheral venous line.

Incubator Enclosed cot for the baby with controlled temperature, humidity and oxygen.

Intubation A procedure whereby a a tube is inserted into the trachea to help a patient's breathing (also called endotracheal intubation).

Mean BP Mean blood pressure as measured via the arterial catheter.

Neopuff Provision of inflationary breaths using a bag via a mask or endotracheal tube connector.

NICU Neonatal Intensive Care Unit.

Peripheral temperature (T2) Temperature measured at the periphery of the body, typically at the toe.

Peripheral venous line Narrow tube inserted into a vein in a limb.

Phototherapy Treatment involving the exposure of the skin to UV light.

Re-intubation Procedure of changing an endotracheal tube.

Re-site probes/sensors Moving a probe or sensor to another location on the baby.

SaO2 Oxygen saturation in the blood as measured by pulse oximetry.

SpO2 Pulse oximeter sensor.

Suction Removal of secretions from the oro/naso pharyngeal area and or from an endotracheal tube.

TcPCO2 Partial pressure of carbon dioxide in the blood as measured by the transcutaneous sensor.

TcPO2 Partial pressure of oxygen in the blood as measured by the transcutaneous sensor.

Transcutaneous sensor Sensor on the baby's skin for measuring TcPO2 and TcPCO2.

Ventilation Respiratory support for babies who are unable or too immature to breathe independently.

References

[1] S. Afantenosa, V. Karkaletsisa, and P. Stamatopoulos. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

[2] W. Aigner, S. Miksch, W. Mueller, H. Schumann, and C. Tominski. Visualizing time-oriented data: A systematic view. *Computers and Graphics*, 31(3):401–409, 2007.

[3] E. Alberdi, J.-C. Becher, K. J. Gilhooly, J. R. W. Hunter, R. H. Logie, A. Lyon, N. McIntosh, and J. Reiss. Expertise and the interpretation of computerised physiological data: Implications for the design of computerised physiological monitoring in neonatal intensive care. *International Journal of Human Computer Studies*, 55(3):191–216, 2001.

[4] N. Aleks, S. Russell, M. Madden, K. Staudenmayer, M. Cohen, D. Morabito, and G. Manley. Probabilistic detection of short events, with application to critical care monitoring. In *Advances in Neural Information Processing Systems 21*. MIT Press, 2009.

[5] E. André. The generation of multimedia presentations. In R. Dale, H. Moisl, and H. Somers, editors, *A Handbook of*

Natural Language Processing: Techniques and applications for the processing of language as text, pages 305–327. Marcel Dekker Inc., 2000.

[6] A. Aris, B. Schneiderman, C. Plaisant, G. Shmueli, and W. Jonk. Representing unevenly-spaced time series data for visualization and interactive exploration. In *Proceedings of the 10th International Conference on Human-Computer Interaction (INTERACT-05)*, 2005.

[7] S. Badaloni and M. Giacomin. The algebra IA$^{\text{fuz}}$: A framework for qualitative fuzzy temporal reasoning. *Artificial Intelligence*, 170(10):872–908, 2006.

[8] J. A. Bateman. Kpml: The kometpenman (multilingual) development environment. Technical Report 0.8, Institut fur Integrierte Publikations- und Informationssysteme (IPSI), GMD, Darmstadt, 1995.

[9] A. Belz. And now with feeling: Developments in emotional language generation. Technical Report ITRI-03-21, Information Technology Research Institute, University of Brighton, 2003.

[10] A. Belz and E. Kow. System building cost vs. output quality in data-to-text generation. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-09)*, 2009.

[11] A. Belz and E. Reiter. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 2006.

[12] D. Bental and A. Cawsey. Personalized and adaptive systems for medical consumer applications. *Communications of the ACM*, 45(5):62–63, 2002.

[13] B. Bohnet, F. Lareau, and L. Wanner. Automatic production of multilingual environmental information. In *Proceedings of the 21st Conference on Informatics for Environmental Protection (EnviroInfo-07)*, 2007.

[14] B. Buchanan and E. Shortliffe. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Boston, Ma., 1984.

[15] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. Interactive pattern search in time series. In *Proceedings of the Conference on Visualization and Data Analysis (VDA-05)*, 2005.

[16] M. Caenepeel. Aspect and text structure. *Linguistics*, 33:213–253, 1995.

[17] A. Cawsey, R. Jones, and J. Pearson. An evaluation of a personalised health information system for patients with cancer. *User Modelling and User-Adapted Interaction*, 10:47–72, 2000.

[18] M.-C. Chambrin. Alarms in the intensive care unit: How can the number of false alarms be reduced? *Critical Care*, 5(4):184–188, 2001.

[19] L. Chittaro. Information visualization and its application to medicine. *Artificial Intelligence in Medicine*, 22:81–88, 2001.

[20] L. Chittaro and C. Combi. Temporal granularity and indeterminacy in reasoning about actions and change: An approach based on the event calculus. *Annals of Mathematical Artificial Intelligence*, 36(1–2):81–119, 2002.

[21] J. Cimino, H. Min, and Y. Perl. Consistency across the hierarchies of the UMLS semantic network and metathesaurus. *Journal of Biomedical Informatics*, 36:450–461, 2003.

[22] Clevermed Limited. The Bagder system, 2007.

[23] J. Coch. Interactive generation and knowledge administration in MULTIMETEO. In *Proceedings of the 9th International Workshop on Natural Language Generation (IWNLG-98)*, 1998.

[24] S. Cunningham, S. Deere, A. Symon, R. A. Elton, and N. . McIntosh. A randomized, controlled trial of computerized physiologic trend monitoring in an intensive care unit. *Critical Care Medicine*, 26:2053–2059, 1998.

[25] R. Dale. StockReporter. Available at: http://www.ics.mq.edu.au/lt-gdemo/StockReporter, 2003.

[26] F. de Rosis and F. Grasso. Affective natural language generation. In A. Paiva, editor, *Affective Interactions*, Lecture Notes in AI, pages 204 – 218. Springer, 2000.

[27] F. de Rosis, F. Grasso, and D. Berry. Refining instructional text generation after evaluation. *Artificial Intelligence in Medicine*, 17(1):1–36, 1999.

[28] R. DeMier, M. Hynan, R. Hatfield, M. Varner, H. Harris, and R. Manniello. A measurement model of perinatal stressors: Identifying risk for postnatal emotional distress in mothers of high-risk infants. *Journal of Clinical Psychology*, 56:89–100, 2000.

[29] M. Dojat, F. Pachet, Z. Guessoum, D. Touchard, A. Harf, and L. Brochard. Néoganesh: A working system for the automated control of assisted ventilation in ICUs. *Artificial Intelligence in Medicine*, 11(2):97–117, 1997.

[30] D. Dowty. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Springer, Berlin, 1979.

[31] D. Dubois, H. Allel, and H. Prade. Fuzziness and uncertainty in temporal reasoning. *Journal of Universal Computer Science*, 9(9):1168–1194, 2003.

[32] R. Dunbar and M. Spoors. Social networks, support cliques, and kinship. *Human Nature*, 6(3):273–290, 1995.

[33] M. Elhadad and J. Robin. An overview of SURGE: A reusable comprehensive syntactic realization component. In *Proceedings of the 8th International Workshop on Natural Language Generation (INLG-96)*, 1996.

[34] L. S. Elting and G. P. Body. Is a picture worth a thousand medical words? a randomized trial of reporting formats for medical research data. *Methods of Information in Medicine*, 30:145–150, 1991.

[35] L. S. Elting, C. G. Martin, S. B. Cantor, and E. B. Rubenstein. Influence of data display formats on physician investigators' decisions to stop clinical trials: Prospective trial with repeated measures. *British Medical Journal*, 318:1527–1531, 1999.

[36] G. Ewing, Y. Freer, R. Logie, J. Hunter, N. McIntosh, S. Rudkin, and L. Ferguson. Role and experience determine decision support interface requirements in a neonatal intensive care environment. *Journal of Biomedical Informatics*, 36:240–249, 2003.

[37] L. Ferres, A. Parush, S. Roberts, and G. Lindgaard. Helping people with visual impairments gain access to graphical information through natural language: The igraph system. In *Proceedings of the 10th International Conference on Computers Helping People with Special Needs (ICCHP-06)*, 2006.

[38] Y. Freer, L. Ferguson, G. Ewing, J. Hunter, R. Logie, S. Rudkin, and N. McIntosh. Mismatched concepts in a neonatal intensive care unit (NICU): Further issues for computer decision support? *Journal of Clinical Monitoring and Computing*, 17:441–447, 2002.

[39] Y. Freer, A. Lyon, B. Stenson, and C. Coyle. BabyLink: Improving communication among clinicians and with parents with babies in intensive care. *British Journal of Healthcare computing and information Management*, 22(2):34–36, 2005.

[40] E. Friedman-Hill. *Jess in Action: Java Rule-based Systems*. Manning Publications Co, USA, 2003.

[41] F. Gao, S. Sripada, J. Hunter, and F. Portet. Using temporal constraints to integrate signal analysis and domain knowledge in medical event detection. In *Proceedings of the Twelfth European Conference on Artificial Intelligence in Medicine (AIME-09)*, 2009.

[42] A. Garg, N. Adhikari, H. McDonald, M. Rosas-Arellano, P. Devereaux, J. Beyene, J. Sam, and R. Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *Journal of the American Medical Association*, 293(10):1223–1238, 2005.

[43] A. Gatt and F. Portet. Text content and task performance in the evaluation of a natural language generation system. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2009.

[44] A. Gatt and E. Reiter. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-09)*, 2009.

[45] P. Gervás, B. L. Loenneker-Rodman, J. Mester, and F. Peinado. Narrative models: Narratology meets artificial intelligence. In *Proceedings of the Workshop on Computational Models of Literary Analysis, 5th International Conference on Language Resources and Evaluation (LREC-06)*, 2006.

[46] E. Goldberg, N. Driedger, and R. I. Kittredge. Using natural language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53, 1994.

[47] A. Graesser, M. Singer, and T. Trabasso. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371–395, 1994.

[48] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

[49] C. Hallett, R. Power, and D. Scott. Summarisation and visualisation of e-health data repositories. In *Proceedings of the UK E-Science All-Hands Meeting*, 2006.

[50] S. Hanks and D. Madigan. Probabilistic temporal reasoning. In M. Fisher, D. Gabbay, and L. Vila, editors, *Handbook of Temporal Reasoning in Artificial Intelligence*, pages 239–261. Elsevier, Amsterdam, Netherlands, 2005.

[51] M. Harris. Building a large-scale commercial NLG system for an EMR. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG-08)*, 2008.

[52] Health Power House. The Euro Health Consumer Index 2008. Electronic, accessed December 2008, 2008.

[53] J. Hitzeman. *Temporal Adverbials and the Syntax-Semantics Interface.* PhD thesis, University of Rochester, 1993.

[54] W. Horn, S. Miksch, G. Egghart, C. Popow, and F. Paky. Effective data validation of high-frequency data: time-point, time-interval, and trend-based methods. *Computers in Biology and Medicine*, 27(5):389–409, 1997.

[55] E. Hovy. *Generating Natural Language Under Pragmatic Constraints*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[56] D. Hueske-Kraus. Suregen-2: A shell system for the generation of clinical documents. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, 2003.

[57] D. Hueske-Kraus. Text generation in clinical medicine: A review. *Methods of Information in Medicine*, 42(1):51–60, 2003.

[58] B. L. Humphreys and D. A. Lindberg. The UMLS project: Making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81(2):170–177, 1993.

[59] J. Hunter. Tsnet: A distributed architecture for time series analysis. In *Proceedings of the Workshop on Intelligent Data Analysis in Biomedecine and Pharmacology (IDAMAP-06)*, 2006.

[60] J. Hunter, G. Ewing, L. Ferguson, Y. Freer, R. Logie, P. McCue, and N. McIntosh. The NEONATE database. In *Proceedings of the AIME-03 Workshop on Intelligent Data Analysis in Medicine and Pharmacology and Knowledge-Based Information Management in Anaesthesia and Intensive Care*, 2003.

[61] L. Iordanskaja, M. Kim, R. Kittredge, B. Lavoie, and A. Polguere. Generation of extended bilingual statistical reports. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*, 1992.

[62] P. W. Jordan. Contextual influences on attribute selection for repeated descriptions. In K. van Deemter and R. Kibble, editors, *Information Sharing: Reference and Presupposition in Natural Language Generation and Understanding*. CSLI Publications, Stanford, Ca., 2002.

[63] R. Jucks and R. Bromme. Choice of words in doctor-patient communication: An analysis of health-related internet sites. *Health Communication*, 21(3):267–277, 2007.

[64] M. G. Kahn, L. M. Fagan, and L. B. Sheiner. Combining physiologic models and symbolic methods to interpret time-varying patient data. *Methods of Information in Medicine*, 30(3):167–178, 1991.

[65] A. Karasimos and A. Isard. Multi-lingual evaluation of a natural language generation system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 2004.

[66] E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings of the IEEE International Conference on Data Mining (ICDM-01)*, 2001.

[67] R. Kibble. Cb or not Cb? Centering Theory applied to NLG. In *Proceedings of the ACL-99 Workshop on Discourse and Reference Structure*, 1999.

[68] R. Kibble and R. Power. An integrated framework for text planning and pronominalisation. In *Proceedgins of the 1st International Conference on Natural Language Generation (INLG-00)*, 2000.

[69] R. Klabunde. Lexical choice of modal expressions. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG-07)*, 2007.

[70] R. Kosara and S. Miksch. Visualization methods for data analysis and planning in medical applications. *International Journal of Medical Informatics*, 68:141–153, 2002.

[71] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CSLI Publications, Stanford, 2002.

[72] K. Kukich. Design of a knowledge-based report generator. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL-83)*, 1983.

[73] W. Labov. *Language in the Inner City*. University of Pennsylvania Press, Pennsylvania, 1971.

[74] W. Labov. Uncovering the event structure of narrative. In D. Tannen and J. E. Alatis, editors, *Linguistics, Language and the Real World: Discourse and Beyond*. Georgetown University Press, Washington, D.C., 2001.

[75] A. S. Law, Y. Freer, J. Hunter, R. H. Logie, N. McIntosh, and J. Quinn. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19(3):183–194, 2005.

[76] I. Mani, M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovksy. Machine learning of temporal relations. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-06)*, 2006.

[77] S. B. Martins, Y. Shahar, D. Goren-Bar, M. Galperin, H. Kaizer, L. V. Basso, D. McNaughton, and M. K. Goldstein. Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data. *Artificial Intelligence in Medicine*, 43(1):17–34, 2008.

[78] K. F. McCoy and M. Strube. Taking time to structure discourse: Pronoun generation beyond accessibility. In *Proceedings of the 1999 Meeting of the Cognitive Science Society (CogSc-99)*, 1999.

[79] N. McIntosh, A. J. Lyon, J. Reiss, J. C. Becher, R. Logie, K. Gilhooley, E. Alberdi, and J. Hunter. The cognitive processes of doctors and nurses in the interpretation of physiological monitoring data in the neonate. *Early Human Development*, 58(1):73, 2000.

[80] K. McKeown, J. Robin, and M. Tanenblatt. Tailoring lexical choice to the user's vocabulary in multimedia explanation generation. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, 1993.

[81] C. Mellish, D. Scott, L. Cahill, D. Paiva, R. Evans, and M. Reape. A reference architecture for Natural Language Generation systems. *Natural Language Engineering*, 12(1):1–34, 2006.

[82] M. Moens and M. Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28, 1988.

[83] W. Moncur, J. Masthoff, and E. Reiter. What do you want to know? Investigating the information requirements of patient supporters. In *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS-08)*, 2008.

[84] W. Mueller and H. Schumann. Visualization methods for time-dependent data: An overview. In *Proceedings of the 35th Winter Simulation Conference*, 2003.

[85] N. F. Noy, M. Crubezy, R. W. Fergerson, H. Knublauch, S. W. Tu, J. Vendetti, and M. A. Musen. Protege-2000: An open-source ontology-development and knowledge-acquisition environment. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA-00)*, 2003.

[86] J. Oberlander and A. Lascarides. Preventing false implicatures: Interactive defaults for text generation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*, 1992.

[87] E. Ochs. Narrative. In T. van Dijk, editor, *Discourse as Structure and Process*. Sage Publications, UK, 1997.

[88] M. O'Donnel, C. Mellish, J. Oberlander, and A. Knott. Ilex: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3):225–250, 2001.

[89] J. Palma, J. Juareza, M. Camposa, and R. Marina. Fuzzy theory approach for temporal model-based diagnosis: An application to medical domains. *Artificial Intelligence in Medicine*, 38(2):197–218, 2006.

[90] I. Paraboni, K. van Deemter, and J. Masthoff. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 32(2):229–254, 2007.

[91] C. Paris. Tailoring object descriptions to the user's level of expertise. *Computational Linguistics*, 14(3):64–78, 1988.

[92] R. Passonneau. Situations and intervals. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL-87)*, 1987.

[93] N. Peek, M. Verduijn, E. de Jonge, and B. de Mol. An empirical comparison of four procedures for filtering monitoring data. In *Proceedings of the Workshop on Intelligent Data Analysis in Biomedicine and Pharmacology*, 2007.

[94] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Conference on Advanced Visual Interfaces (AVI-04)*, 2004.

[95] F. Portet and A. Gatt. Towards a possibility-theoretic approach to uncertainty in medical data interpretation for text generation. In *Proceedings of the Workshop on Knowledge Representation for Healthcare (KR4HC-2009)*, 2009.

[96] F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7–8):789–816, 2009.

[97] S. Powsner and E. Tufte. Graphical summary of patient status. *The Lancet*, 344:386–389., 1994.

[98] F. Puppe, M. Atzmueller, G. Buscher, M. Huettig, H. Luehrs, and H. P. Buscher. Application and evaluation of a medical knowledge-system in sonography (sonoconsult). In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-08)*, 2008.

[99] J. Quinn, C. Williams, and N. McIntosh. Factorial switching Kalman Filters applied to condition monitoring in neonatal intensive care. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[100] H. Reichenbach. *Elements of Symbolic Logic*. Macmillan, New York, 1947/1966.

[101] E. Reiter. An architecture for data-to-text systems. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG-07)*, 2007.

[102] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, 2000.

[103] E. Reiter, A. Gatt, F. Portet, and M. van der Meulen. The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG-08)*, 2008.

[104] E. Reiter, R. Robertson, and L. Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58, 2003.

[105] E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169, 2005.

[106] E. Reiter, R. Turner, N. Alm, R. Black, M. Dempster, and A. Waller. Using NLG to help language-impaired users tell stories and participate in social dialogues. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-09)*, 2009.

[107] A. Roberts, R. Gaizauskas, M. Hepple, and Y. Guo. Mining clinical relationships from patient narratives. *BMC Bioinformatics*, 9(Supp.11, S3), 2008.

[108] V. Ryabov and A. Trudel. Probabilistic temporal interval networks. In *11th International Symposium on Temporal Representation and Reasoning (TIME'04)*, 2004.

[109] B. Schneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, 1996.

[110] B. Schneiderman and B. Bederson. Maintaining concentration to achieve task completion. In *Proceedings of the Conference on Designing for User Experience (DUX-05)*, 2005.

[111] C. Smith. The domain of tense. In J. Guéron and J. Lecarme, editors, *The Syntax of Time*. MIT Press, Cambridge, Ma., 2003.

[112] S. Sripada, E. Reiter, and I. Davy. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10, 2003.

[113] M. Stacey and C. McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial Intelligence in Medicine*, 39(1):1–24, 2007.

[114] O. Stock, M. Zancanaro, P. Busetta, C. Callaway, A. Krueger, M. Kruppa, T. Kuflik, E. Not, and C. Rocchi. Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User Modeling and User-Adapted Interaction*, 17(3):257–304, 2007.

[115] B. Strople and P. Ottani. Can technology improve intershift report? what the research reveals. *Journal of Professional Nursing*, 22(3):197–204, 2006.

[116] K. Tan, P. Dear, and S. Newell. Clinical decision support systems for neonatal care. *Cochrane Database of Systematic Reviews*, 2, 2005.

[117] F. Tehrani and J. Roum. Intelligent decision support systems for mechanical ventilation. *Artificial Intelligence in Medicine*, 44(3):171–182, 2008.

[118] P. Terenziani, R. Snodgrass, A. Bottrighi, M. Torchio, and G. Molino. Extending temporal databases to deal with telic/atelic medical data. In *Proceedings of the 10th Conference on Artificial Intelligence in Medicine (AIME 05)*, 2005.

[119] R. Turner, S. Sripada, E. Reiter, and I. Davy. Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. In *Proceedings of the Conference on Applications and Innovations in Intelligent Systems XV*, 2007.

[120] K. van Deemter. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222, 2006.

[121] K. van Deemter, B. Krenn, P. Piwek, M. Schroeder, M. Klesen, and S. Baumann. Fully generated scripted dialogue for embodied conversational agents. *Artificial Intelligence*, 172(10):1219–1244, 2008.

[122] M. van der Meulen, R. H. Logie, Y. Freer, C. Sykes, N. McIntosh, and J. Hunter. When a graph is poorer than 100 words: A comparison of computerised Natural Language Generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, to appear.

[123] I. van der Sluis and C. Mellish. Towards empirical evaluation of affective tactical NLG. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-09)*, 2009.

[124] Z. Vendler. Verbs and times. *The Philosophical Review*, 66(2):143–160, 1957.

[125] B. L. Webber. The interpretation of tense in discourse. In *Proceedings of the 25th Meeting of the Association for Computational Linguistics (ACL-87)*, 1987.

[126] M. White, R. Rajkumar, and S. Martin. Towards broad coverage surface realization with CCG. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, 2007.

[127] S. Williams and E. Reiter. Generating basic skills reports for low-skilled readers. *Journal of Natural Language Engineering*, 14(4):495–525, 2008.

[128] W. D. Winn. Contributions of perceptual and cognitive processes to the comprehension of graphics. In W. Schnotz and R. Kulhavy, editors, *Comprehension of Graphics*, pages 3–27. Elsevier, Amsterdam, 1994.

[129] J. Yu, E. Reiter, J. Hunter, and C. Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13:25–49, 2007.

[130] L. Zhou and G. Hripcsaka. Temporal reasoning with medical dataa review with emphasis on medicalnatural language processing. *Journal of Biomedical Informatics*, 40(2):183–202, April 2007.

[131] R. Zwaan. Time in narrative comprehension: A cognitive perspective. In D. H. Schram and G. J. Steen, editors, *Psychology and Sociology of Literature*, pages 71–86. John Benjamins, Amsterdam, 2001.

[132] R. Zwaan. Time in language, situation models and mental simulations. *Language Learning*, 58:13–26, 2008.

[133] R. Zwaan and G. A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185, 1998.